Introduction to Data Science
Supermarket Sales

Introduction

Objective

This dataset of sales observations aims to predict customer satisfaction levels for two groups: members and normal visitors. Satisfaction is categorized into two levels: low (0) and high (1). Key influencing factors include Quantity, Total, Gross Income, and Unit Price. The goal is to use Rating as the predicted variable, as customer satisfaction helps identify potential revenue generators.

Data Source

The dataset of 1000 supermarket, Supermarket_sales.xlsx, was sourced from Kaggle to analyze customer satisfaction trends. Categorizing the rating variable provides clear insights into overall satisfaction, aiding in sustainable growth analysis and developing effective marketing strategies.

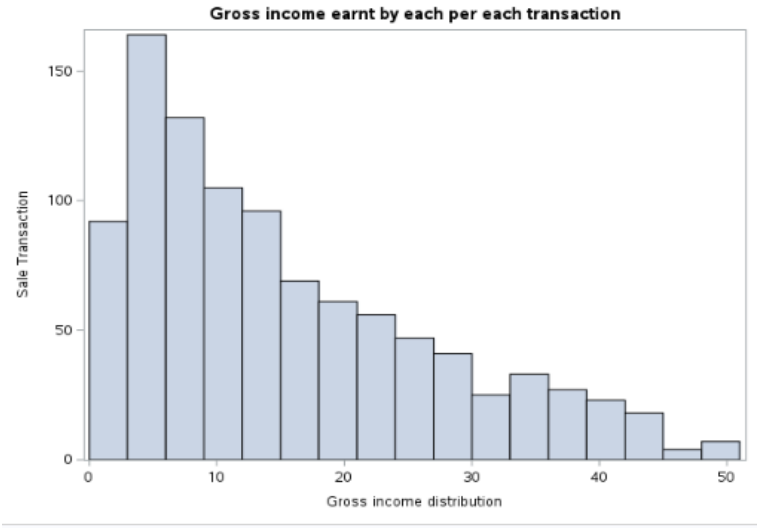
Sno	Variable	Description
1	Customer_Type	Member, Normal
2	Gender	Male, Female
3	Product_line	Product_line,Health and beauty,Electronic accessories,Home and lifestyle,Sports and travel,Food and beverages,Fashion accessories
4	Unit_Price	Price of a single unit purchased
5	Quantity	Number of items purchased
6	Tax	Tax charged on the purchase, calculated as (Unit Price * Quantity) * 5%
7	Total	Total revenue for a transaction, calculated as (Unit Price * Quantity) + Tax
8	Date	Date of purchase
9	Time	Time of purchase
10	Payment	Mode of Payment
11	Cogs	Cost of goods purchased
12	Gross_income	Total revenue minus COGS (Total - Cogs)

13	Rating	0 (Unsatisfied), 1 (Satisfied)

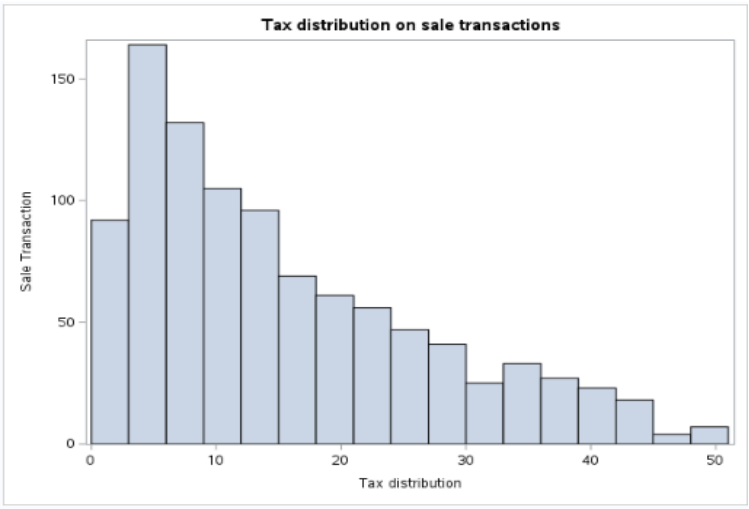
Data Partition

Customers with satisfaction scores between 0-7 are categorized as low-level, while scores above 7 are high-level. The data is split 80-20 for training and testing, ensuring a balanced dataset without the need for over-sampling or under-sampling.

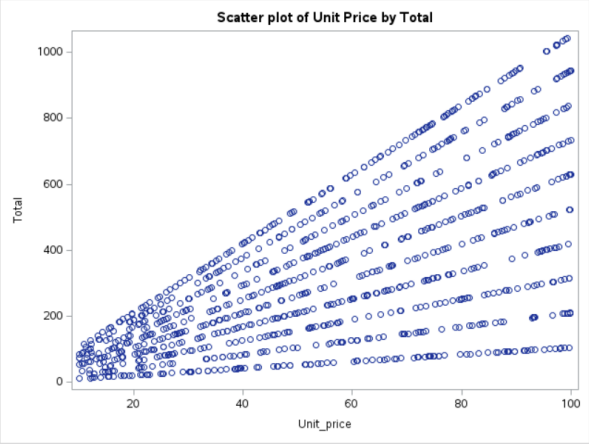
Graphs and Summary Statistics

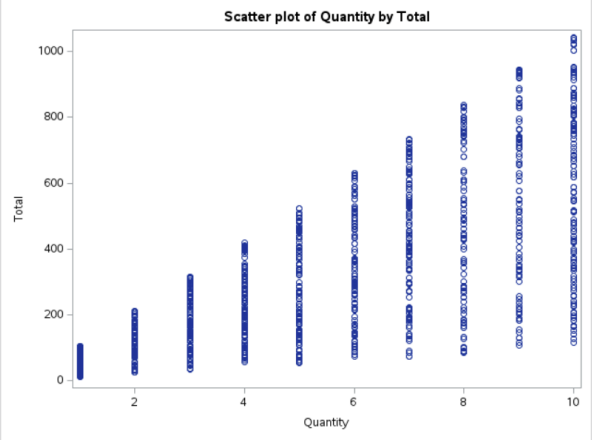


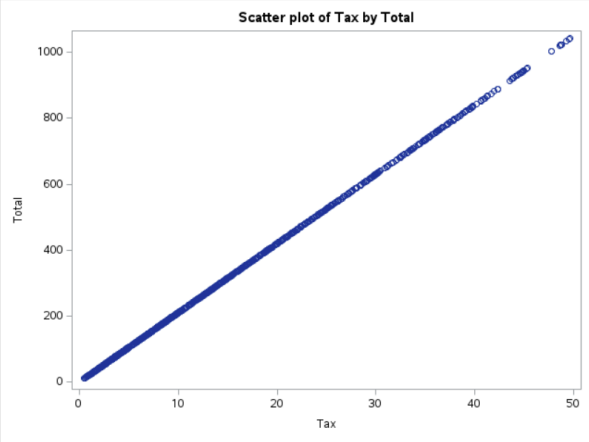
The histogram shows the distribution of sales transactions by gross income. It is right-skewed, indicating that high-gross-income transactions are rare. Most transactions fall within the lowest category, generating a gross income between 0 and 10.

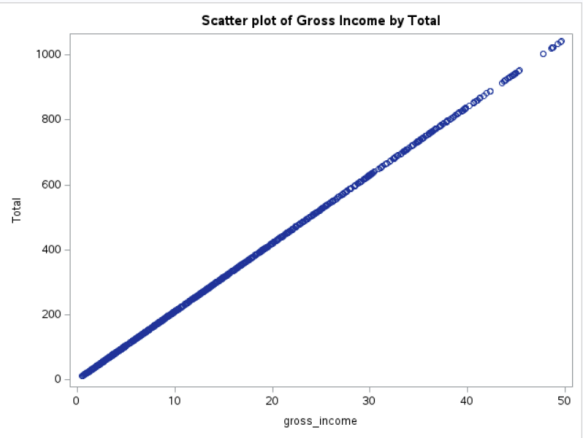


The tax distribution histogram mirrors the income distribution, showing a right skew. This suggests that most transactions have low tax amounts, with higher tax transactions being rare. The majority of transactions fall within the 0 to 10 tax range.

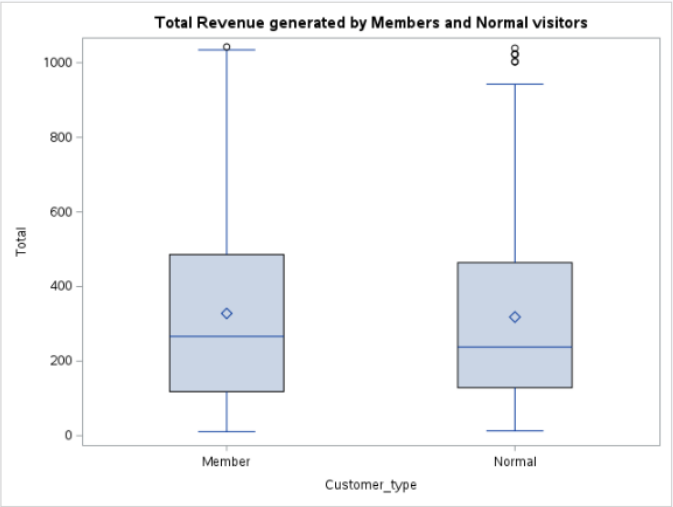




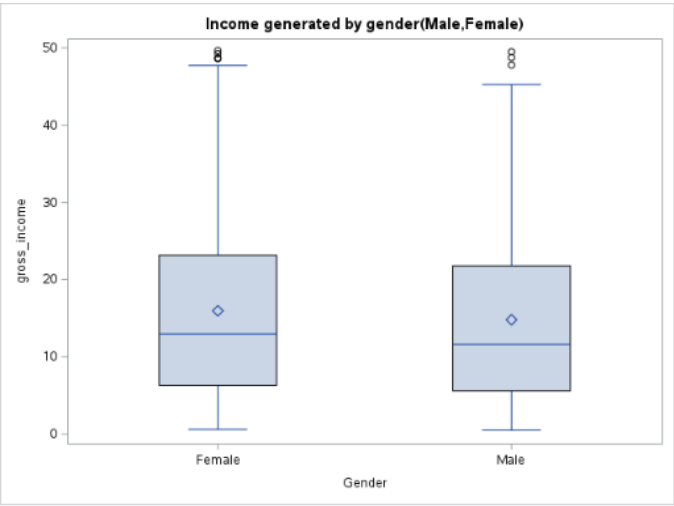




The scatter plots reveal a clear linear relationship between Total and other variables, such as Unit Price, Tax, Gross Income, and Quantity. As these variables increase, Total also rises. This suggests that higher values for Unit Price, Tax, Quantity, and Gross Income are positively correlated with an increased Total. Given the strong correlation observed between Total and the other variables in the plots, we chose Total as the primary variable for our models. Since the calculation of Total inherently includes contributions from Unit Price, Quantity, Tax, and Gross Income, incorporating Total alone is sufficient for our analysis.

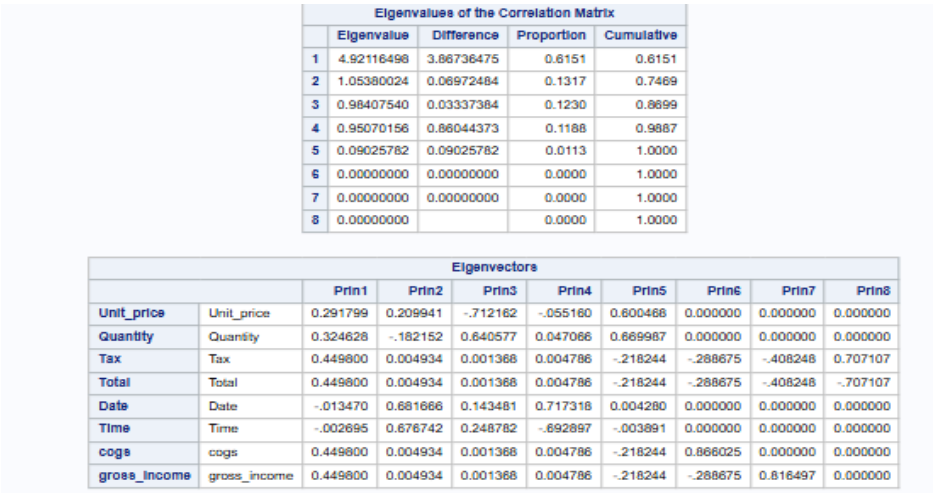


Both member and normal customers show similar mean and median revenues, with members generating slightly higher revenue. Both categories have right-skewed distributions, with members displaying a slightly higher revenue trend. The left whiskers are nearly identical, and the interquartile ranges (IQRs) indicate similar revenue distributions. Notably, normal visitors have a higher deviation due to more outliers compared to members.



The plot shows that female customers generate slightly more revenue than male customers. Both categories have right-skewed distributions, with the whiskers extending upward. Additionally, the IQR, median, mean, and outliers are similar for both male and female customers.

Principal Component Analysis



Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.92116498	3.86736475	0.6151	0.6151
2	1.05380024	0.06972484	0.1317	0.7469
3	0.98407540	0.03337384	0.1230	0.8699
4	0.95070156	0.86044373	0.1188	0.9887
5	0.09025782	0.09025782	0.0113	1.0000
6	0.00000000	0.00000000	0.0000	1.0000
7	0.00000000	0.00000000	0.0000	1.0000
8	0.00000000		0.0000	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Unit_price	Unit_price	0.291799	0.209941	-.712162	-.055160	0.600468	0.000000	0.000000	0.000000
Quantity	Quantity	0.324628	-.182152	0.640577	0.047066	0.669987	0.000000	0.000000	0.000000
Tax	Tax	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	-.408248	0.707107
Total	Total	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	-.408248	-.707107
Date	Date	-.013470	0.681666	0.143481	0.717318	0.004280	0.000000	0.000000	0.000000
Time	Time	-.002695	0.676742	0.248782	-.692897	-.003891	0.000000	0.000000	0.000000
cogs	cogs	0.449800	0.004934	0.001368	0.004786	-.218244	0.866025	0.000000	0.000000
gross_income	gross_income	0.449800	0.004934	0.001368	0.004786	-.218244	-.288675	0.816497	0.000000

The PCA analysis suggests that correlation is a better approach, as the variables are measured in different units. Regarding the eigenvalues, the first principal component explains 61% of the data variability, while the second accounts for 13%. Together, the first three principal components explain 86% of the total variability.

The most important features in each principal component are as follows:

- **Prin1:** Tax, Total, COGS, and Gross Income
- **Prin2:** Date and Time
- **Prin3:** Quantity
- **Prin4:** Date and Time
- **Prin5:** Quantity
- **Prin6:** COGS
- **Prin7:** Gross Income
- **Prin8:** Total and Tax

Model Selection

We have evaluated and compared the performance of several predictive models, including Logistic Regression, CART, Neural Networks, and Discriminant Analysis.

The dataset is split into an 80% training set and a 20% validation set, comprising 800 records for model building. The dependent variable is the rating, with a focus on predicting **Rating 1** as the event of interest.

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	SALES
Random Number Seed	12345
Sampling Rate	0.8
Sample Size	800
Selection Probability	0.8
Sampling Weight	0
Output Data Set	SALES_PART

Model 1: Logistic Regression

Dummy variables are created for categorical variables such as Customer_Type, Gender, Product_Line, and Payment to fit the logistic regression model appropriately.

The LOGISTIC Procedure

Model Information		
Data Set	WORK.SALES_TRAIN	
Response Variable	Rating	Rating
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	800
Number of Observations Used	800

Response Profile		
Ordered Value	Rating	Total Frequency
1	0	412
2	1	388

Probability modeled is Rating='1'.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Is_Member	0.986	0.745	1.304
Is_Female	1.066	0.804	1.413
Is_Product_line_HB	1.122	0.686	1.834
Is_Product_line_EA	0.779	0.487	1.246
Is_Product_line_HL	0.840	0.522	1.350
Is_Product_line_ST	0.789	0.490	1.273
Is_Product_line_FB	1.156	0.722	1.852
Total	1.000	0.999	1.001
Date	0.999	0.993	1.004
Time	1.000	1.000	1.000
Is_Payment_Ewallet	1.072	0.766	1.499
Is_Payment_CC	1.137	0.804	1.609

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	24.3634	61.5348	0.1568	0.6922
Is_Member	1	-0.0143	0.1427	0.0100	0.9203
Is_Female	1	0.0636	0.1438	0.1957	0.6582
Is_Product_line_HB	1	0.1147	0.2511	0.2086	0.6478
Is_Product_line_EA	1	-0.2499	0.2397	1.0869	0.2972
Is_Product_line_HL	1	-0.1745	0.2423	0.5185	0.4715
Is_Product_line_ST	1	-0.2364	0.2438	0.9404	0.3322
Is_Product_line_FB	1	0.1453	0.2402	0.3657	0.5453
Total	1	0.000058	0.000285	0.0409	0.8398
Date	1	-0.00111	0.00285	0.1520	0.6966
Time	1	-8.34E-6	6.201E-6	1.8085	0.1787
Is_Payment_Ewallet	1	0.0692	0.1714	0.1628	0.6866
Is_Payment_CC	1	0.1287	0.1770	0.5283	0.4673
cogs	0	0	.	.	.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.6143	12	0.8145
Score	7.5871	12	0.8165
Wald	7.5298	12	0.8207

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	55.7	Somers' D	0.115
Percent Discordant	44.3	Gamma	0.115
Percent Tied	0.0	Tau-a	0.057
Pairs	159856	c	0.557

Fit Statistics for SCORE Data										
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC
WORK.SALES_TRAIN	800	-550.4	0.4625	1126.701	1127.164	1187.601	1187.601	0.009473	0.012634	0.557289
WORK.SALES_VALID	200	-138.5	0.4750	303.0507	305.0077	345.9288	345.9288	-0.00387	-0.00517	0.516732

The FREQ Procedure

Frequency
Row Pct

Table of Rating by I_Rating			
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	269 65.29	143 34.71	412
1	227 58.51	161 41.49	388
Total	496	304	800

The FREQ Procedure

Frequency
Row Pct

Table of Rating by I_Rating			
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	61 57.01	46 42.99	107
1	49 52.69	44 47.31	93
Total	110	90	200

Stepwise Selection Methods

Forward: Only the intercept is included, as none of the variables tested for inclusion meet the statistical significance criterion at the 0.05 level.
Backward: Similar results, with the elimination of the Total, Gross Income, and COGS variables.

Stepwise: Only the intercept is included, as no variables tested for inclusion meet the statistical significance criterion at the 0.05 level.

Logistic Regression Final Model

The final model was built using the variables suggested by the backward selection method, ensuring that all dummy variables of a variable are retained. All measures in the final model are identical to those in the initial model.

	Training				Validation			
	AUC	Error Rate	Sensitivity	Specificity	AUC	Error Rate	Sensitivity	Specificity
Logistic Regression Model	0.56	46%	42%	65%	0.52	48%	47%	57%
Logistic Regression Final Model	0.56	46%	42%	65%	0.52	48%	47%	57%

Evaluation of Logistic regression model

- Likelihood Ratio is 0.81 (p-value > 0.05), indicating the model is not statistically significant.
- Variable Significance, Is_Member, Total, Date, Time p-values less than 0.05, suggesting they are statistically significant predictors in this model.
- AUC (Area Under the Curve) is 0.56, which is below the desired threshold of 1. This indicates below-average performance, though within an acceptable range for initial analysis.

Performance Metrics

- Error Rate: High error rates for both model fit and accuracy indicate poor classification performance.
- Sensitivity and Specificity:
    - Training Set: Sensitivity: 42%, Specificity: 65%
    - Validation Set: Sensitivity improved to 4%, while specificity decreased to 57%.

Conclusion

The model demonstrates a high error rate in both its fit and accuracy, indicating poor classification performance. The Area Under the Curve (AUC) remains consistent at 0.56 across both models, reflecting below-average predictive capability.

Notably, the AIC and BIC values are significantly lower in the model accuracy assessment compared to the initial fit, suggesting slight improvements. Sensitivity and specificity metrics further highlight the model's limitations: on the training set, sensitivity is 42% and specificity is 65%, while validation results show a modest increase in sensitivity to 47% and a decrease in specificity to 57%.

There is no evidence of overfitting, but overall, the model’s performance falls below expectations.

Model 2: CART

CART (Classification and Regression Trees) model, which is well-suited for our objective of handling a classification event. Given this classification focus, the model was developed using two key measures, the Gini Index and Entropy.

GINI

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	0
Number of Leaves Before Pruning	142
Number of Leaves After Pruning	1
Model Event Level	1

Number of Observations Read	1000
Number of Observations Used	1000
Number of Training Observations Used	800
Number of Validation Observations Used	200

ENTROPY

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	10
Number of Leaves Before Pruning	122
Number of Leaves After Pruning	85
Model Event Level	1

Number of Observations Read	1000
Number of Observations Used	1000
Number of Training Observations Used	800
Number of Validation Observations Used	200

Model	Training				Validation			
	AUC	Error Rates	Sensitivity	Specificity	AUC	Error Rates	Sensitivity	Specificity
Gini	0.5	48.50%	0%	100%	0.5	48.50%	0%	100%
Entropy	0.89	21.25%	85%	73%	0.55	45.50%	62%	47.66%

The Entropy model produced a larger and more complex tree, resulting in 85 leaves after pruning. This complexity allowed it to capture more intricate patterns within the data. In contrast, the Gini model generated an overly simplistic tree with only one leaf, failing to capture any meaningful patterns.

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Training	0	300	112	0.2718
	1	58	330	0.1495
Validation	0	51	56	0.5234
	1	35	58	0.3763

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	85	0.1334	0.2125	0.8505	0.7282	0.5667	0.2669	213.5	0.8906
Validation	85	0.3350	0.4550	0.6237	0.4766	0.7280	0.3535	134.0	0.5503

Performance Metrics

- Error Rate: High error rates for model fit is 0.2125, whereas for the model accuracy is increased to 0.4550 indicating poor classification performance.
- Sensitivity and Specificity:
    - Training Set: Sensitivity: 85.05%, Specificity: 72.82%
    - Validation Set: Sensitivity: 62.37%, Specificity: 47.66%

Model Fit Indicators

- AUC on the training data is 0.8906 and, on the test, data is 0.5503. The significant drop in AUC from training to test data suggests overfitting.

Conclusion

The sensitivity and specificity are high on the training set, but drop significantly on the validation set, suggesting poor model performance on test data. The AUC is 0.8906 on the training data but drops to 0.5503 on the test data, showing overfitting.

Model 3: Neural Networks

Models	Hidden Layers	Neuron in each Layer	Training			Valdiation		
			Error Rate	Sensitivity	Specificity	Error Rate	Sensitivity	Specificity
1	1	8	45%	47.68%	61.89%	43%	52.69%	60.75%
2	1	20	45%	49.74%	59.22%	44%	52.69%	58.88%
3	1	25	44%	44.59%	66.99%	48%	44.09%	58.88%
4	1	30	45%	45.10%	64.08%	43%	51.61%	61.68%
5	1	35	45%	42.78%	65.53%	5%	38.71%	58.88%
6	1	40	44%	46.13%	64.32%	48%	43.01%	59.81%
7	**3**	**8**	**41%**	**36.60%**	**79.61%**	**46%**	**30.11%**	**74.77%**

Performance after Adding 3 hidden layer with 8 neurons in each

Train: Misclassification Rate	Valid: Misclassification Rate
0.4125	0.4600

Performance Metrics

Error Rate: The misclassification rate has reduced for training (0.41) and slight change in validation sets (0.46) as compared to previous model.

Sensitivity and Specificity:

Training Set: Sensitivity: 36.60%, Specificity: 79.61%

Validation Set: Sensitivity: 30.11%, Specificity: 74.77%

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	328 79.61	84 20.39	412
1	246 63.40	142 36.60	388
Total	574	226	800

The FREQ Procedure

Frequency Row Pct	Table of Rating by I_Rating		
	I_Rating(Into: Rating)		
Rating(Rating)	0	1	Total
0	80 74.77	27 25.23	107
1	65 69.89	28 30.11	93
Total	145	55	200

In Model 7, compared to Model 1, the sensitivity decreased in both the training and validation sets. However, the specificity improved in both sets. Therefore, adding 3 hidden layers with the same number of neurons enhanced specificity but reduced sensitivity, resulting in a moderate

Thus, we are choosing Model 7 as final model in neural networks.

Model 4: Discriminant Analysis

Class Level Information					
Rating	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	412	412.0000	0.515000	0.515000
1	1	388	388.0000	0.485000	0.485000

412 observations belong to class 0, ratings unsatisfactory. 388 observations belong to class 1, where ratings are satisfactory. Corresponds to 52% and 48%.

Discriminant Analysis	Training			Valdiation		
	Error Rate	Sensitivity	Specificity	Error Rate	Sensitivity	Specificity
Without Prior	48%	14.69%	86.41%	49%	8.60%	90.65%
With Prior(60-40)	40%	100%	0%	40%	100%	0%

Number of Observations and Percent Classified into Rating			
From Rating	0	1	Total
0	356 86.41	56 13.59	412 100.00
1	331 85.31	57 14.69	388 100.00
Total	687 85.88	113 14.13	800 100.00
Priors	0.515	0.485	

Number of Observations and Percent Classified into Rating			
From Rating	0	1	Total
0	97 90.65	10 9.35	107 100.00
1	85 91.40	8 8.60	93 100.00
Total	182 91.00	18 9.00	200 100.00
Priors	0.515	0.485	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.1359	0.8531	0.4838
Priors	0.5150	0.4850	

Error Count Estimates for Rating			
	0	1	Total
Rate	0.0935	0.9140	0.4914
Priors	0.5150	0.4850	

Conclusion

The model performed poorly in terms of sensitivity but showed strong results for specificity. After applying the prior proportion, the model's performance declined, as it struggled to accurately identify specificity.

Model Comparison and Evaluation

	Training				Validation			
Model	**AUC**	**Error Rate**	**Sensitivity**	**Specificity**	**AUC**	**Error Rate**	**Sensitivity**	**Specificity**
Logistic Regression Model	0.56	46%	42%	65%	0.52	48%	47%	57%
CART Model(Entropy)	**0.89**	**21%**	**85%**	**73%**	**0.55**	**46%**	**62%**	**48%**
Neural Network (3 hidden layer with 8 Neurons each)		41%	37%	80%		46%	30%	75%
Discriminant Analysis		48%	15%	86%		49%	9%	91%

The CART model performs well on the training data, achieving 85% sensitivity, 73% specificity, and the lowest error rate of 21%. It also performed reasonably on the validation data, with 62% sensitivity, 48% specificity, and a 46% error rate.

While the model demonstrates a strong ability to fit the training data, making it a suitable choice for predicting customer satisfaction, the weaker performance on the validation set and the higher error rate suggest signs of overfitting.

Conclusion

Practical Significance and Implications

- Targeted Marketing: Insights on spending behaviors and satisfaction levels of any customer that can be used to create targeted promotions for specific customers including those who are members and those who are not.

- Operational Efficiency: Analysis of date and time trends might be useful for the proper staffing of employees and the proper supply chain management during peak hours.

Inferences

- Key factors influencing customer satisfaction through data-driven analysis - Date and Time.
- Models can predict satisfaction trends across similar customer demographics.

Shortcomings

- Potential biases in the data collection process could impact the generalizability of results.
- The dataset size is relatively small, which might limit the robustness of the findings.

Suggestions for Improvement

- Increasing the dataset size and incorporating the latest data, as we currently have data from 2019, will contribute to greater diversity and help capture a broader range of customer behaviors.
- Incorporate additional variables, such as customer reviews, to improve the accuracy and relevance of predictions.

References

https://www.kaggle.com/code/aryantiwari123/supermarket-sales-prediction/inpu

SAS Code

Group8_Project_Final_Code.sas

```
/* Import */

proc import out=supermarket datafile="/home/u63814242/sasuser.v94/Data Science/Project/supermarket_sales.xlsx"
dbms=xlsx replace;
run;

/* Plots */

/*histogram*/
proc sgplot data=supermarket;
   histogram gross_income/scale=count;
   title "Gross income earnt by each per each transaction";
   xaxis label="Gross income distribution";
    yaxis label="Sale Transaction";
run;

proc sgplot data=supermarket;
   histogram Tax/scale=count;
   title "Tax distribution on sale transactions";
   xaxis label="Tax distribution";
    yaxis label="Sale Transaction";
run;

/*Side by Side boxplot*/
proc sgplot data=supermarket;
     vbox Total /category=Customer_type;
     title "Total Revenue generated by Members and Normal visitors";
run;

proc sgplot data=supermarket;
     vbox gross_income /category=gender;
     title "Income generated by gender(Male,Female)";
run;
```

```
/*Scatter Plot*/
proc sgplot data=supermarket;
     scatter x=unit_price y=gross_income;
run;

proc sgplot data=supermarket;
     scatter x=quantity y=gross_income;
run;

proc sgplot data=supermarket;
     scatter x=tax y=gross_income;
run;

proc sgplot data=supermarket;
     scatter x=total y=gross_income;
run;

 /*numerical summary(descriptive statistics)*/
proc means data=supermarket n mean median skew maxdec=2;
var gross_income Tax ;
run;

/*numerical summary(descriptive statistics)*/
proc means data=supermarket n mean median skew maxdec=2;
var total cogs ;
run;


/*Principal component analysis*/
data supermarket1;
    set supermarket;
    drop Rating;
run;

proc princomp data=supermarket1 out=supermarket2;
run;

/* logistic regression */
data supermarket;
    set supermarket;
    if Customer_type='Member' then Is_Member=1; else Is_Member=0;
    if Gender='Female' then Is_Female=1; else Is_Female=0;
    if Payment='Ewallet'  then Is_Payment_Ewallet=1; else Is_Payment_Ewallet=0;
    if Payment='Credit card'  then Is_Payment_CC=1; else Is_Payment_CC=0;
    if Product_line='Health and beauty'  then Is_Product_line_HB=1; else Is_Product_line_HB=0;
    if Product_line='Electronic accessories'  then Is_Product_line_EA=1; else Is_Product_line_EA=0;
    if Product_line='Home and lifestyle'  then Is_Product_line_HL=1; else Is_Product_line_HL=0;
    if Product_line='Sports and travel'  then Is_Product_line_ST=1; else Is_Product_line_ST=0;
    if Product_line='Food and beverages'  then Is_Product_line_FB=1; else Is_Product_line_FB=0;
run;

proc surveyselect data=supermarket samprate=.8 method=srs outall out=supermarket_Part seed=12345;
run;

data supermarket_train supermarket_valid;
    set supermarket_Part;
    if selected=1 then output supermarket_train; else output supermarket_valid;
run;
```

```
proc logistic data=supermarket_train outmodel=model_logistic;
    model Rating (event="1")= Is_Member Is_Female Is_Product_line_HB Is_Product_line_EA
                              Is_Product_line_HL Is_Product_line_ST Is_Product_line_FB
                              Total Date Time   Is_Payment_Ewallet
                              Is_Payment_CC cogs;
run;

proc logistic inmodel=model_logistic;
    score data=supermarket_train fitstat out=supermarket_trainOut;
    score data=supermarket_valid fitstat out=supermarket_validOut;
run;

proc freq data=supermarket_trainOut;
    table Rating*i_Rating /nopercent nocol;
run;

proc freq data=supermarket_validOut;
    table Rating*i_Rating/nopercent nocol;
run;


/* Selection method */
proc logistic data=supermarket_train outmodel=model1;
    model Rating (event="1")= Is_Member Is_Female Is_Product_line_HB Is_Product_line_EA
                              Is_Product_line_HL Is_Product_line_ST Is_Product_line_FB
                              Total Date Time   Is_Payment_Ewallet
                              Is_Payment_CC cogs/selection=forward;
run;/*only intercept*/


proc logistic data=supermarket_train outmodel=model1;
    model Rating (event="1")= Is_Member Is_Female Is_Product_line_HB Is_Product_line_EA
                              Is_Product_line_HL Is_Product_line_ST Is_Product_line_FB
                              Total Date Time   Is_Payment_Ewallet
                              Is_Payment_CC cogs/selection=backward;
run;/*few variables eliminated*/

proc logistic data=supermarket_train outmodel=model1;
    model Rating (event="1")= Is_Member Is_Female Is_Product_line_HB Is_Product_line_EA
                              Is_Product_line_HL Is_Product_line_ST Is_Product_line_FB
                              Total Date Time   Is_Payment_Ewallet
                              Is_Payment_CC cogs/selection=stepwise;
run;/*only intercept*/

/*Final logistic Model*/
proc logistic data=supermarket_train outmodel=model_logistic_final;
    model Rating (event="1")= Is_Member Is_Female Is_Product_line_HB Is_Product_line_EA
                              Is_Product_line_HL Is_Product_line_ST Is_Product_line_FB
                              Total Date Time   Is_Payment_Ewallet
                              Is_Payment_CC;
run;

proc logistic inmodel=model_logistic_final;
    score data=supermarket_train fitstat out=supermarket_trainOut_final;
    score data=supermarket_valid fitstat out=supermarket_validOut_final;
run;

proc freq data=supermarket_trainOut_final;
    table Rating*i_Rating /nopercent nocol;
run;

proc freq data=supermarket_validOut_final;
    table Rating*i_Rating/nopercent nocol;
run;
```

```
/*CART*/

proc hpsplit data=supermarket_part nodes=detail;
    partition rolevar=selected(train='1' validate='0');
    class Customer_type Gender Product_line Payment Rating;
    model Rating(event="1")= Customer_type Gender Product_line Total Date Time Payment cogs;
    grow gini;
    prune cc;
run;

proc hpsplit data=supermarket_part nodes=detail;
    partition rolevar=selected(train='1' validate='0');
    class Customer_type Gender Product_line Payment Rating;
    model Rating(event="1")= Customer_type Gender Product_line Total Date Time Payment cogs;
    grow entropy;
    prune cc;
run;


/* Neural Network */

proc hpneural data=supermarket_part;
    partition rolevar=selected(train=1);
    target Rating/level=nom;
    input Customer_type Gender Product_line Payment/level=nom;
    input Total Date Time cogs/level=int;
    hidden 8;
    train maxiter=1000 numtries=10;
    id Rating selected;
    score out=supermarket_out1;
run;

proc freq data=supermarket_out1;
    table Rating*i_Rating/nopercent nocol;
    where selected=1;
run;

proc freq data=supermarket_out1;
    table Rating*i_Rating/nopercent nocol;
    where selected=0;
run;


proc hpneural data=supermarket_part;
    partition rolevar=selected(train=1);
    target Rating/level=nom;
    input Customer_type Gender Product_line Payment/level=nom;
    input Total Date Time cogs/level=int;
    hidden 8;
    hidden 8;
    hidden 8;
    train maxiter=1000 numtries=10;
    id Rating selected;
    score out=supermarket_out2;
run;

proc freq data=supermarket_out2;
    table Rating*i_Rating/nopercent nocol;
    where selected=1;
run;

proc freq data=supermarket_out2;
    table Rating*i_Rating/nopercent nocol;
    where selected=0;
run;

proc hpneural data=supermarket_part;
    partition rolevar=selected(train=1);
    target Rating/level=nom;
    input Customer_type Gender Product_line Payment/level=nom;
    input Total Date Time cogs/level=int;
    hidden 20;
    train maxiter=1000 numtries=10;
    id Rating selected;
    score out=supermarket_out3;
run;
```

```sas
proc freq data=supermarket_out3;
    table Rating*i_Rating/nopercent nocol;
    where selected=1;
run;

proc freq data=supermarket_out3;
    table Rating*i_Rating/nopercent nocol;
    where selected=0;
run;
proc hpneural data=supermarket_part;
    partition rolevar=selected(train=1);
    target Rating/level=nom;
    input Customer_type Gender Product_line Payment/level=nom;
    input Total Date Time cogs/level=int;
    hidden 40;
    train maxiter=1000 numtries=10;
    id Rating selected;
    score out=supermarket_out4;
run;

proc freq data=supermarket_out4;
    table Rating*i_Rating/nopercent nocol;
    where selected=1;
run;

proc freq data=supermarket_out4;
    table Rating*i_Rating/nopercent nocol;
    where selected=0;
run;


/* Discriminant Analysis */

data supermarket_train supermarket_valid;
    set supermarket_part;
    if selected=1 then output supermarket_train; else output supermarket_valid;
    drop Selected Customer_type Gender Product_line Total Date Time Payment cogs;
run;

proc discrim data=supermarket_train testdata=supermarket_valid;
    class Rating;
    priors proportional;
run;

proc discrim data=supermarket_train testdata=supermarket_valid;
    class Rating;
    priors '1'=.6 '0'=.4;
run;
```