Segmenting the Travelers: A Classification-Based Study on Tour Purchase Behavior

MIS-5560 - Introduction to Data Science

Tara Lehr, Akku Jacob Shaji, Jigyasa Soni, Vyshali Poola

Presented to **Dr. Hongyu Gao**

Introduction

- What ? - Direct marketing scenario for a tour package company
    - Analyzing customer demographic and interaction data to improve marketing efficiency.
- Why ? - To identify which customers are most likely to purchase a tour package.
- How ? - By applying data science techniques:
- ‣ **Classification Trees (CART)**
- ‣ **Logistic Regression**
- ‣ **Neural Networks**
- ‣ **Clustering (K-Means, Hierarchical)**

Aim - uncover patterns in customer behavior, segment the customer base, and provide insights to guide future marketing decisions.

Project Overview

Purpose:

This project aims to classify and segment potential tour package customers using data-driven models. The insights generated will help marketing teams personalize offers and allocate resources more effectively.

Dataset:

5000 customer records containing demographics, engagement history, income, and product interaction details.

Variables Include:

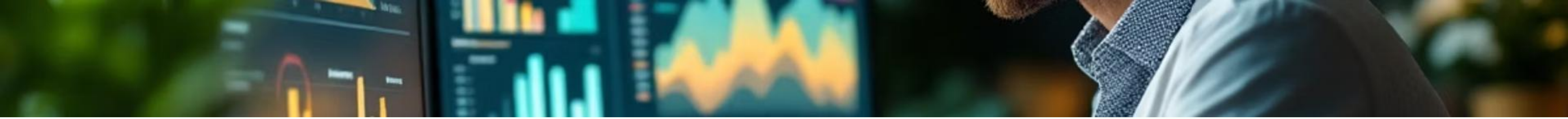
Age, Monthly Income, Gender, Marital Status, Designation, Number of Trips, Duration of Pitch, Product Pitched, Type of Contact, etc.

Target Variable:

`ProdTaken` - a binary indicator showing whether the customer purchased a tour package (`1`) or not (`0`).

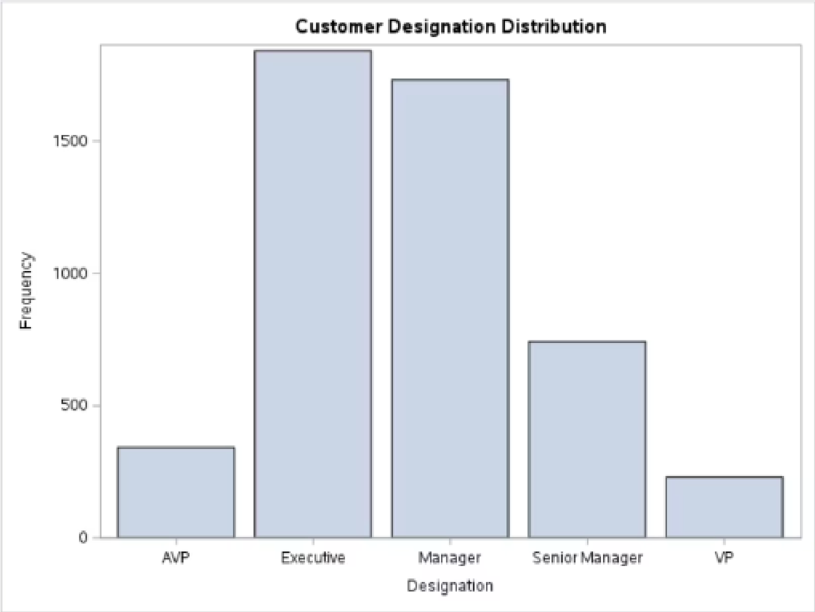
Business Goal:

Identify the profiles of customers most likely to purchase a tour package and build predictive models to improve conversion rates for future marketing campaigns.



Exploratory Data Insights

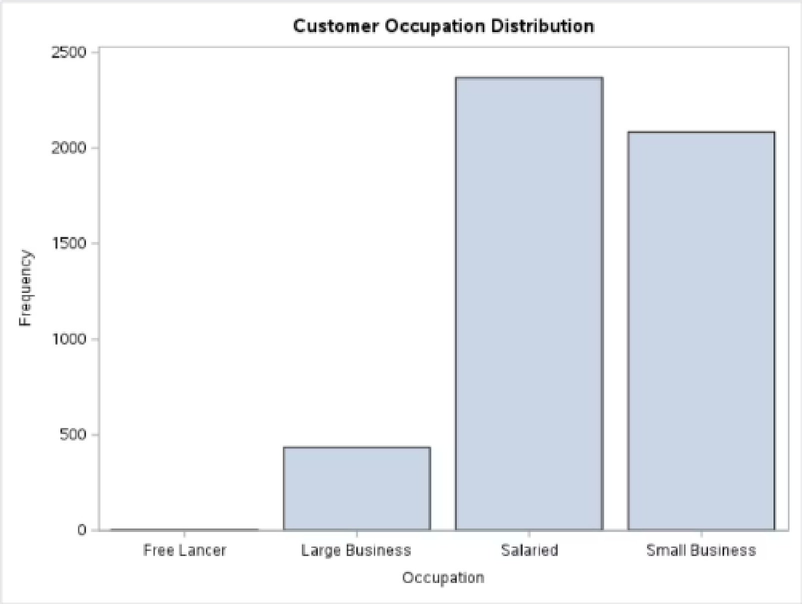
Customer Designation Distribution



Class imbalance in designations justifies

- Dummy encoding
- Stepwise feature selection.

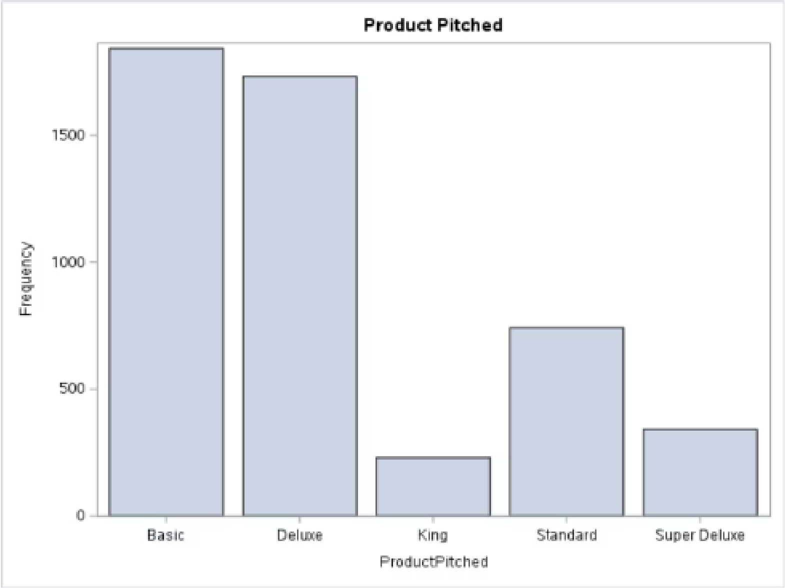
Customer Occupation Distribution



Simplifying the variable improved

- Model stability
- Interpretability.

Product Pitched Distribution



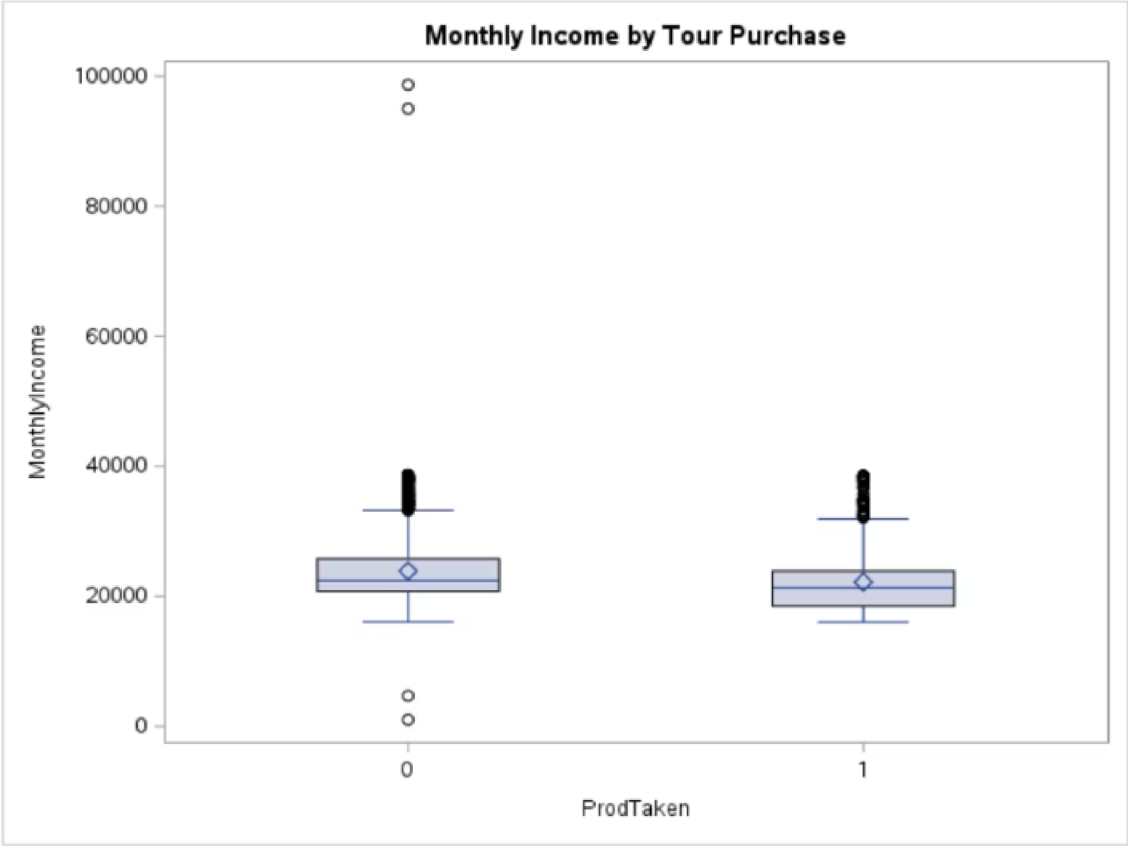
Highlights the need for models that can handle skewed product exposure in predicting conversions.

Exploratory Data Insights

Age by Tour Purchase

Age-based segmentation can help tailor marketing messages for different age brackets.

Monthly Income by Tour Purchase



This helps justify income as a predictor in CART and Logistic Regression

Data Preprocessing

Initial Data Survey

- Reviewed dataset with 4,888 records and 20 variables.
- Generated bar charts & box plots to understand variable distributions

Missing Value Treatment

Identified missing data in key numerical fields:

- `Age` (226), `MonthlyIncome` (233), `DurationOfPitch` (45)
- Replaced missing values using **median imputation** to maintain distribution robustness.

Variable Transformation

- Converted categorical variables like `Designation`, `Gender`, `MaritalStatus`, `Occupation`, `TypeOfContact`, and `ProductPitched` into **binary dummy variables.**
- Resulted in **28 input features** ready for modeling.

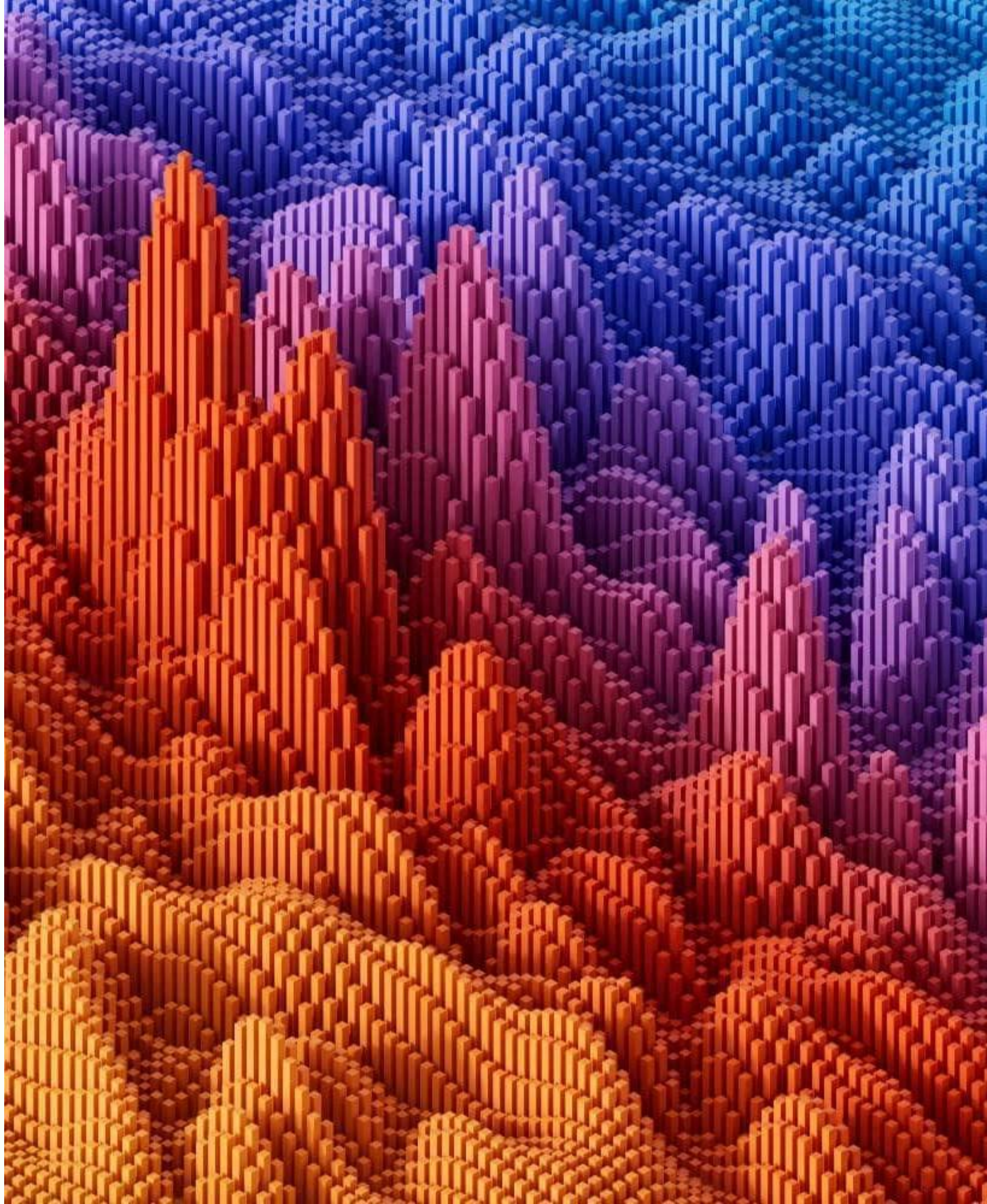
Dataset Partitioning

Split into **60:40** ratio:

- Training: 2,199 records
- Validation: 734 records

Models

- CART (Classification and Regression Tree)
- Logistic Regression
- Neural Networks
- Cluster Analysis



Classification and Regression Tree (CART)

Model Implementation

Two CART models were built using different splitting criteria:

- Model 1: Gini with all variables
- Model 2: Entropy with all variables

Final pruned trees:

- **Gini:** Reduced from **367** to **36 leaves**
- **Entropy:** Reduced from **337** to **43 leaves**

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	30
Maximum Tree Depth Achieved	20
Tree Depth	11
Number of Leaves Before Pruning	367
Number of Leaves After Pruning	36
Model Event Level	0

Number of Observations Read	2933
Number of Observations Used	2933

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	30
Maximum Tree Depth Achieved	22
Tree Depth	10
Number of Leaves Before Pruning	337
Number of Leaves After Pruning	43
Model Event Level	0

Number of Observations Read	2933
Number of Observations Used	2933

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Model Based	0	2281	96	0.0404
	1	251	305	0.4514
Cross Validation	0	2244	133	0.0560
	1	297	259	0.5342

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Model Based	36	0.0986	0.1183	0.9596	0.5486	0.4889	0.1971	578.2	0.8206
Cross Validation	55	0.1199	0.1465	0.9440	0.4658				

Gini Model Insights:

- Consistent AUC across training and validation
- Strong sensitivity → better at identifying positive cases
- Acceptable overfitting risk

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Model Based	0	2290	87	0.0366
	1	250	306	0.4496
Cross Validation	0	2251	126	0.0530
	1	317	239	0.5701

Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Model Based	43	0.0946	0.1149	0.9634	0.5504	0.4681	0.1891	554.7	0.8373
Cross Validation	35	0.1216	0.1511	0.9470	0.4299				

Entropy Model Insights:

- Slightly better AUC and sensitivity
- Higher validation misclassification suggests mild overfitting
- Slightly less balanced compared to Gini

Conclusion:

Entropy model performs slightly better across most metrics, but at the cost of a more complex tree.

Logistic Regression Model -Full

Analysis of Maximum Likelihood Estimates

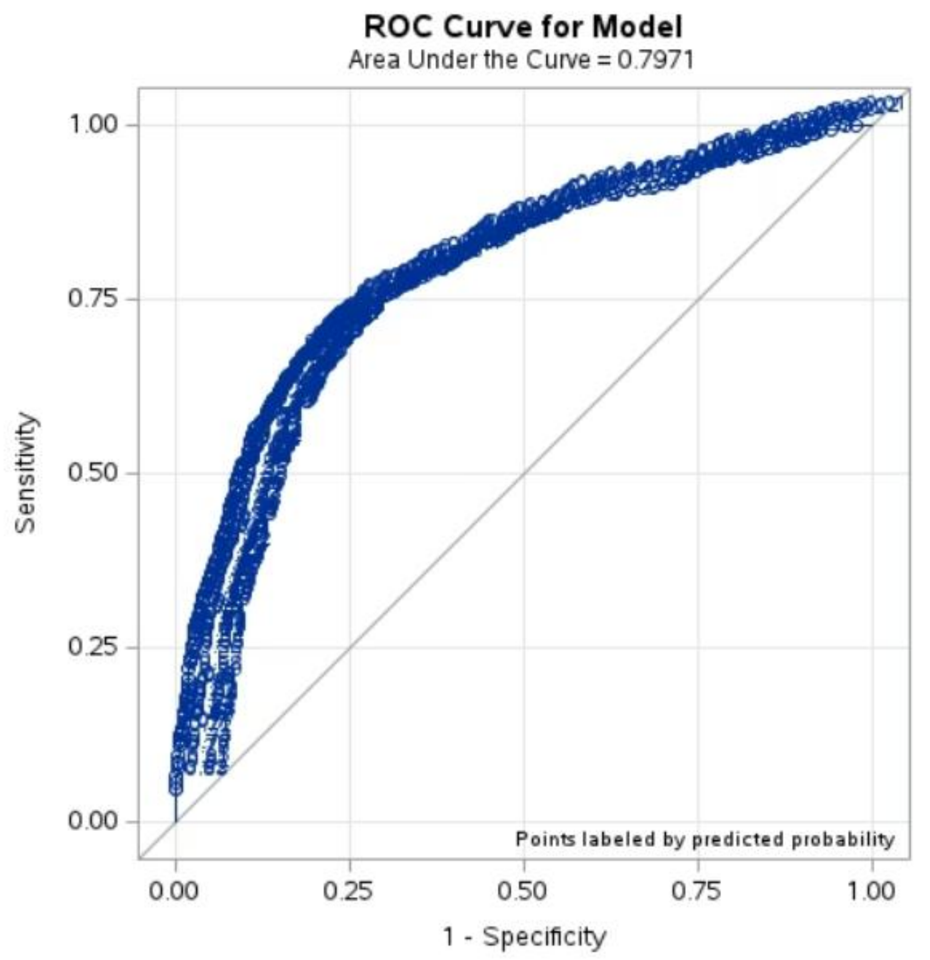
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.7899	227.8	0.0004	0.9832
Age	1	-0.0207	0.00707	8.5463	0.0035
DurationOfPitch	1	0.0354	0.00628	31.8456	<.0001
NumberOfFollowups	1	0.2781	0.0599	21.5583	<.0001
PreferredPropertySta	1	0.4324	0.0646	44.7760	<.0001
NumberOfTrips	1	0.0812	0.0290	7.8452	0.0051
Passport	1	1.5549	0.1104	198.3701	<.0001
OwnCar	1	0.0450	0.1110	0.1644	0.6852
NumberOfChildrenVisi	1	-0.1360	0.0685	3.9455	0.0470
MonthlyIncome	1	4.044E-6	0.000018	0.0489	0.8251
PitchSatisfactionSco	1	0.0943	0.0402	5.5010	0.0190
Designation_AVP	1	-1.2369	0.3496	12.5133	0.0004
Designation_Executiv	1	0.6021	0.2035	8.7565	0.0031
Designation_Manager	1	-0.4316	0.1871	5.3222	0.0211
Designation_SeniorMa	0	0	.	.	.
Designation_VP	1	-0.5373	0.3873	1.9244	0.1654
Gender_Female	1	0.2129	0.3336	0.4073	0.5233
Gender_Male	1	0.5220	0.3259	2.5648	0.1093
MaritalStatus_Divorc	1	-0.7372	0.1935	14.5110	0.0001
MaritalStatus_Marrie	1	-0.7698	0.1664	21.4111	<.0001
MaritalStatus_Single	1	0.4102	0.1827	5.0435	0.0247
MaritalStatus_Unmarr	0	0	.	.	.
Occupation_Large_Bus	1	-9.5265	227.8	0.0017	0.9666
Occupation_Salaried	1	-9.9310	227.8	0.0019	0.9652
Occupation_Small_Bus	1	-9.8392	227.8	0.0019	0.9655
TypeofContact_Compan	1	0.2333	0.8431	0.0766	0.7820
TypeofContact_Self_E	1	-0.1242	0.8397	0.0219	0.8824
ProductPitched_Basic	0	0	.	.	.
ProductPitched_Delux	0	0	.	.	.
ProductPitched_King	0	0	.	.	.
ProductPitched_Stand	0	0	.	.	.
ProductPitched_Super	0	0	.	.	.

Confusion Matrix – Full Model

The FREQ Procedure

Frequency	Table of ProdTaken by predicted_full			
		predicted_full		
	ProdTaken	0	1	Total
	0	1542	49	1591
	1	263	101	364
	Total	1805	150	1955

- Accuracy: 84.04%
- Sensitivity: 27.74%
- Specificity: 96.92%
- AUC: 79.71%



Logistic Regression Model : Stepwise

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Passport		1	1	208.0823		<.0001
2	Designation_Executiv		1	2	131.5513		<.0001
3	MaritalStatus_Single		1	3	55.7425		<.0001
4	PreferredPropertySta		1	4	46.3544		<.0001
5	MaritalStatus_Unmarr		1	5	37.7229		<.0001
6	DurationOfPitch		1	6	33.3284		<.0001
7	NumberOfFollowups		1	7	25.5782		<.0001
8	Designation_AVP		1	8	11.4109		0.0007
9	TypeofContact_Compan		1	9	9.4677		0.0021
10	Gender_Male		1	10	9.0303		0.0027
11	PitchSatisfactionSco		1	11	7.0175		0.0081
12	Age		1	12	6.6280		0.0100
13	NumberOfTrips		1	13	6.8977		0.0086
14	ProductPitched_Stand		1	14	7.5723		0.0059
15	Occupation_Large_Bus		1	15	4.0989		0.0429
16	NumberOfChildrenVisi		1	16	4.0226		0.0449

- Started with all available predictors
- Used entry and stay criteria at $p < 0.05$
- 16 predictors retained

Top predictors by contribution:

- Passport (Chi-Sq: 208.08)
- Designation_Executive (Chi-Sq: 131.55)
- MaritalStatus_Single (Chi-Sq: 55.74)
- PreferredPropertyStar (Chi-Sq: 46.35)

Logistic Regression Model : Stepwise Cont.

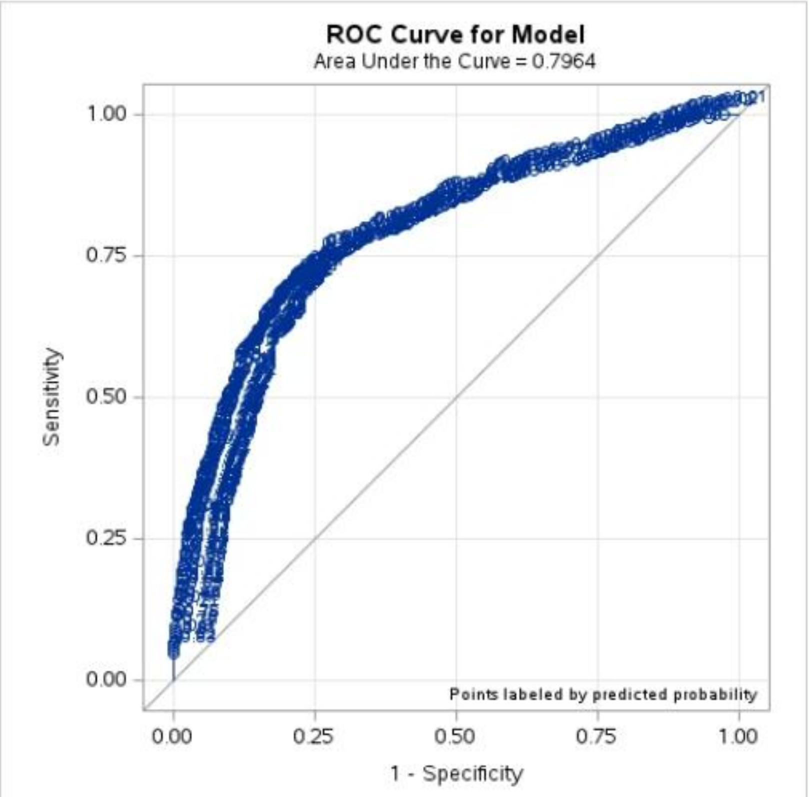
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.1324	0.4755	166.3535	<.0001
Age	1	-0.0208	0.00672	9.6030	0.0019
DurationOfPitch	1	0.0357	0.00625	32.5894	<.0001
NumberOfFollowups	1	0.2842	0.0583	23.7697	<.0001
PreferredPropertySta	1	0.4363	0.0644	45.8709	<.0001
NumberOfTrips	1	0.0804	0.0287	7.8189	0.0052
Passport	1	1.5530	0.1100	199.1697	<.0001
NumberOfChildrenVisi	1	-0.1329	0.0663	4.0149	0.0451
PitchSatisfactionSco	1	0.0952	0.0400	5.6700	0.0173
Designation_AVP	1	-0.7594	0.3181	5.6983	0.0170
Designation_Executiv	1	1.0314	0.1293	63.6798	<.0001
Gender_Male	1	0.3271	0.1107	8.7257	0.0031
MaritalStatus_Single	1	1.1667	0.1297	80.9266	<.0001
MaritalStatus_Unmarr	1	0.7225	0.1485	23.6824	<.0001
Occupation_Large_Bus	1	0.3604	0.1760	4.1931	0.0406
TypeofContact_Compan	1	0.3490	0.1158	9.0770	0.0026
ProductPitched_Stand	1	0.4447	0.1714	6.7322	0.0095

Confusion Matrix – Stepwise Model

The FREQ Procedure

Frequency	Table of ProdTaken by predicted_stepwise			
		predicted_stepwise		
	ProdTaken	0	1	Total
	0	1544	47	1591
	1	262	102	364
	Total	1806	149	1955

- Accuracy: 84.19%
- Sensitivity: 28.02%
- Specificity: 97.05%
- AUC: 79.96%
- Best model for comparisons



Neural Network

Even though Model A had a slightly better AUC, Model B has higher sensitivity and a much stronger odds ratio. That means it's better at actually finding people who will buy, which is what matters most in a marketing campaign. So from a business standpoint, Model B is the better choice for targeting customers more effectively.

Neural Network Model A - 1 layers (with 5 neurons)

Baseline neural network model, Model A. We used one hidden layer with 5 neurons. This structure is simple but powerful enough to capture non-linear patterns in the data.

Number of Observations Read	2933
Number of Observations Used	2933
Number Used for Training	2199
Number Used for Validation	734

Model Information	
Data Source	WORK.TOUR_TRAIN_STD
Architecture	MLP
Number of Input Variables	28
Number of Hidden Layers	1
Number of Hidden Neurons	5
Number of Target Variables	1
Number of Weights	151
Optimization Technique	Limited Memory BFGS

Architecture & Setup:

- 1 hidden layer with 5 neurons
- Optimization: Limited-Memory BFGS
- 50 training iterations
- 28 input variables (dummy-coded)

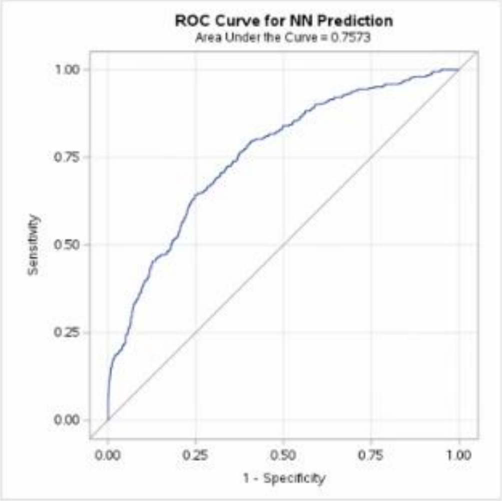
The FREQ Procedure

Frequency	Table of ProdTaken by Prediction		
	Prediction		
ProdTaken	0	1	Total
0	1519	72	1591
1	285	79	364
Total	1804	151	1955

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1881.341	1642.465
SC	1886.919	1653.621
-2 Log L	1879.341	1638.465

Performance Metrics (Validation Set):

- Misclassification Rate: 19.48%
- Sensitivity: 21.7%
- Specificity: 95.5%
- AUC: 0.7573
- Odds Ratio: 64.40



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
NN Prediction	0.7573	0.0139	0.7301	0.7845	0.5147	0.5147	0.1560

Insights:

- Best AUC among all models → good at separating classes
- Moderate sensitivity → detects some purchasers but misses many
- Lower odds ratio → less distinction between buyer and non-buyer scores

Neural Network Model B - 2 layers (8 neurons - 5,3)

Model B increased the complexity–we added another hidden layer and more neurons. This allows the model to learn more detailed patterns in customer behavior.

Model Information	
Data Source	WORK.TOUR_TRAIN_STD
Architecture	MLP
Number of Input Variables	28
Number of Hidden Layers	2
Number of Hidden Neurons	8
Number of Target Variables	1
Number of Weights	167
Optimization Technique	Limited Memory BFGS

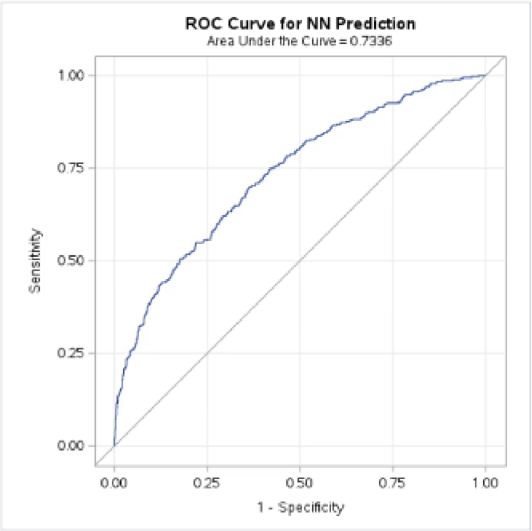
Number of Observations Read	2933
Number of Observations Used	2933
Number Used for Training	2199
Number Used for Validation	734

Confusion Matrix - Neural Network (Tourism)

The FREQ Procedure

Frequency	Table of ProdTaken by Prediction			
		Prediction		
	ProdTaken	0	1	Total
	0	1528	63	1591
	1	276	88	364
	Total	1804	151	1955

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1881.341	1665.367
SC	1886.919	1676.523
-2 Log L	1879.341	1661.367



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
NN Prediction	0.7336	0.0148	0.7047	0.7625	0.4672	0.4672	0.1417

Architecture & Setup:

- 2 hidden layers with 8 neurons each
- Same optimization and input variables as Model A
- Increased model complexity for improved learning

Performance Metrics (Validation Set):

- Misclassification Rate: 20.30%
- Sensitivity: 24.2%
- Specificity: 96.0%
- AUC: 0.7336
- Odds Ratio: 106.75

Insights:

- Higher sensitivity → better at identifying customers likely to buy
- Strong specificity → avoids false positives
- Highest odds ratio → stronger prediction confiden
- Slightly lower AUC but better real-world targeting potential

Unsupervised Learning: Clustering Models

Hierarchical Clustering

Initially explored but produced inconclusive results

K-means Clustering

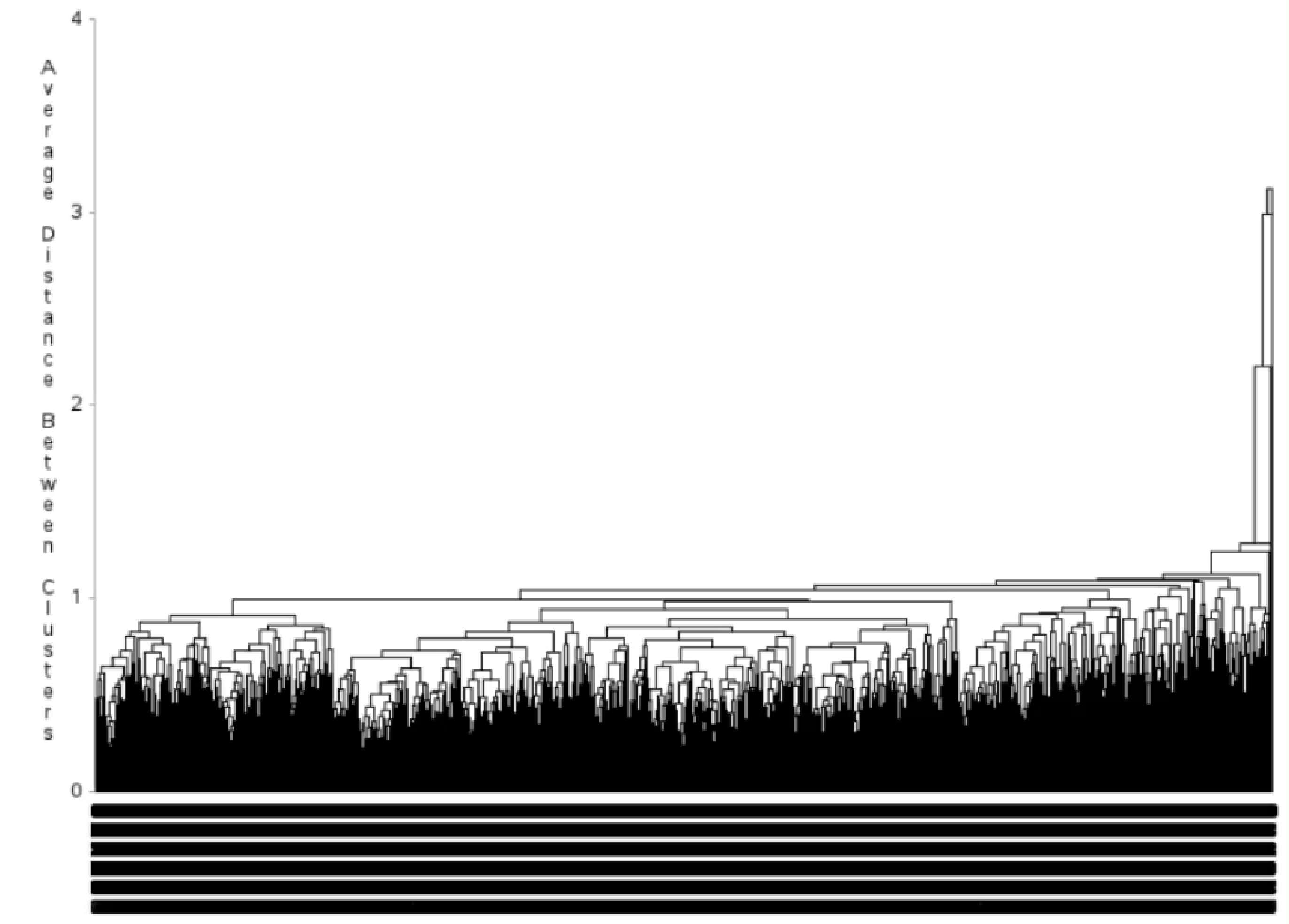
Selected for scalability and flexibility

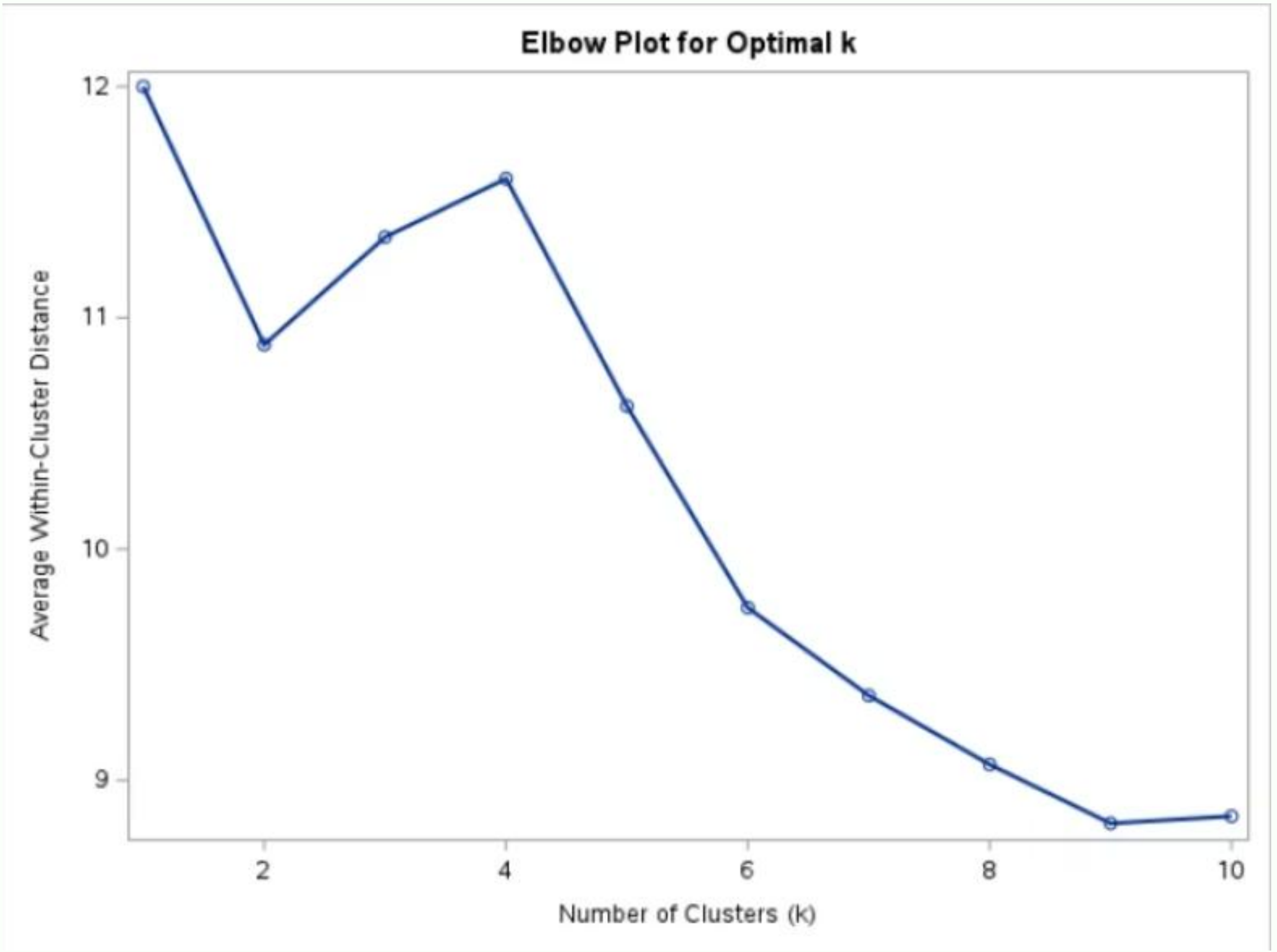
Optimal Cluster Determination

Used elbow method to identify k=6 as optimal

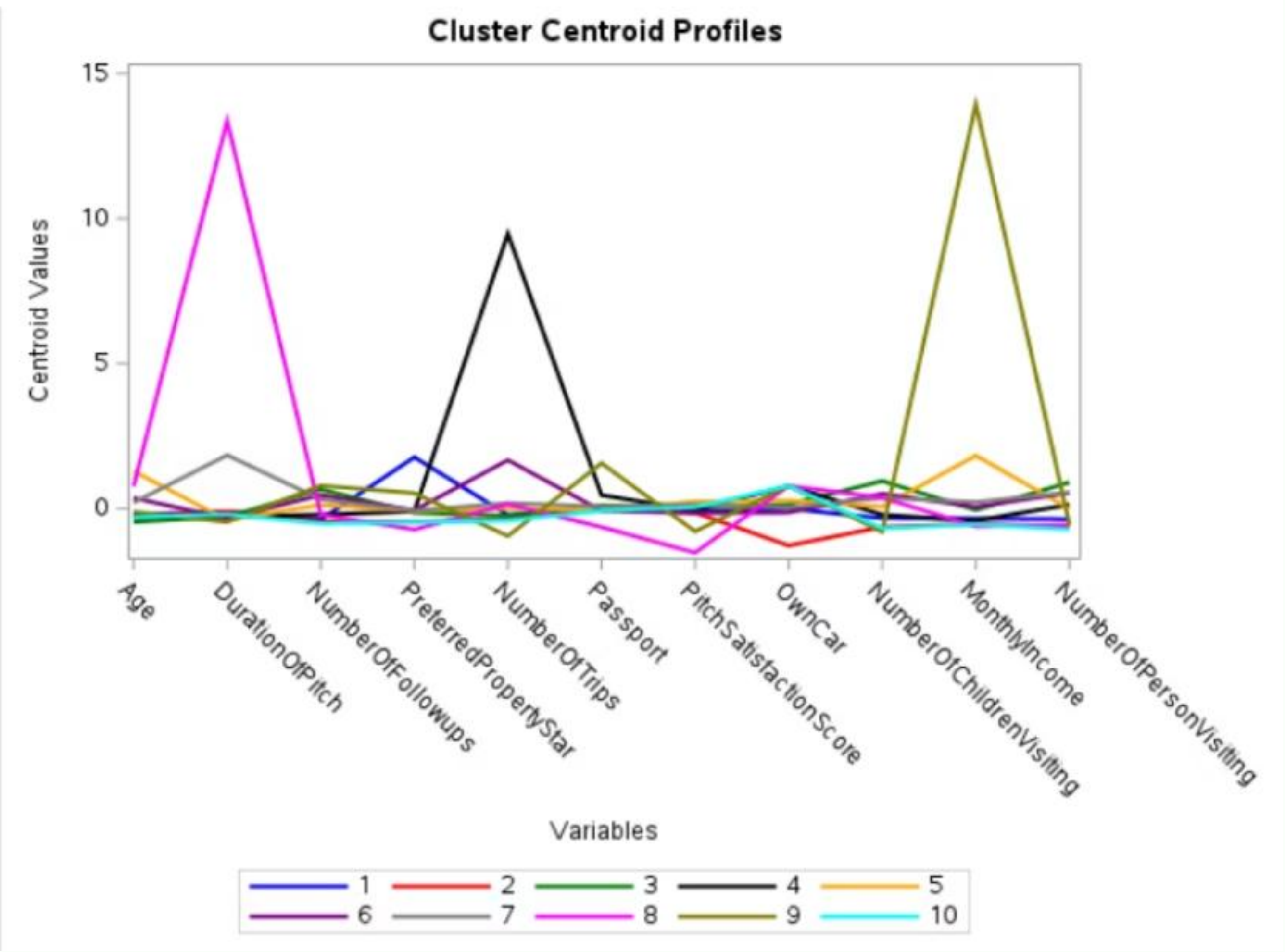
To determine the optimal number of clusters, the FASTCLUS procedure was run iteratively from k=1 to 10, generating metrics such as pseudo F-statistics, observed R-squared, and Cubic Clustering Criterion. A clear inflection occurred at k=6, suggesting this solution offered the best trade-off between performance and interpretability.

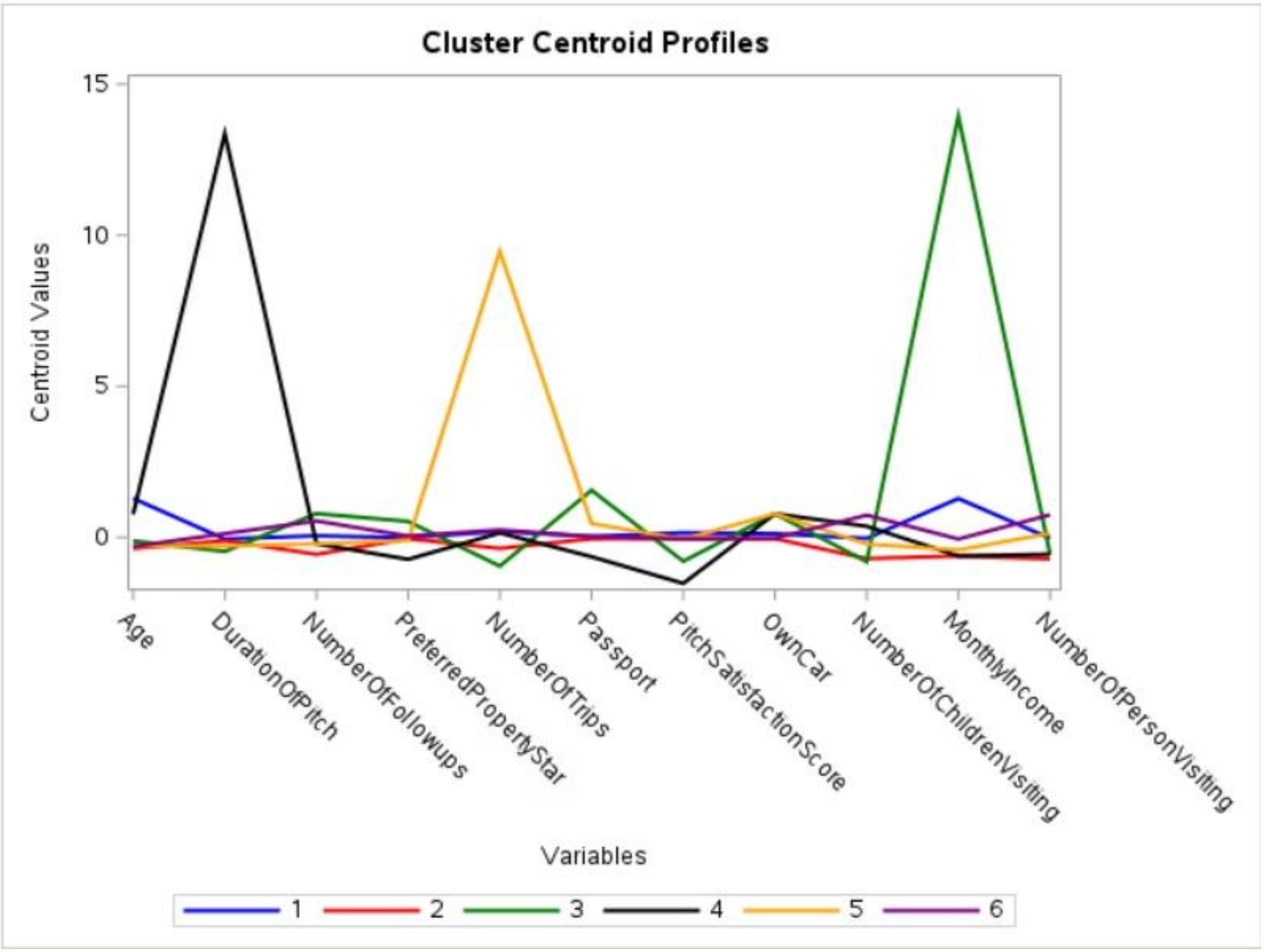
Clustering: Hierarchical vs K-means





Clustering: Hierarchical vs K-means

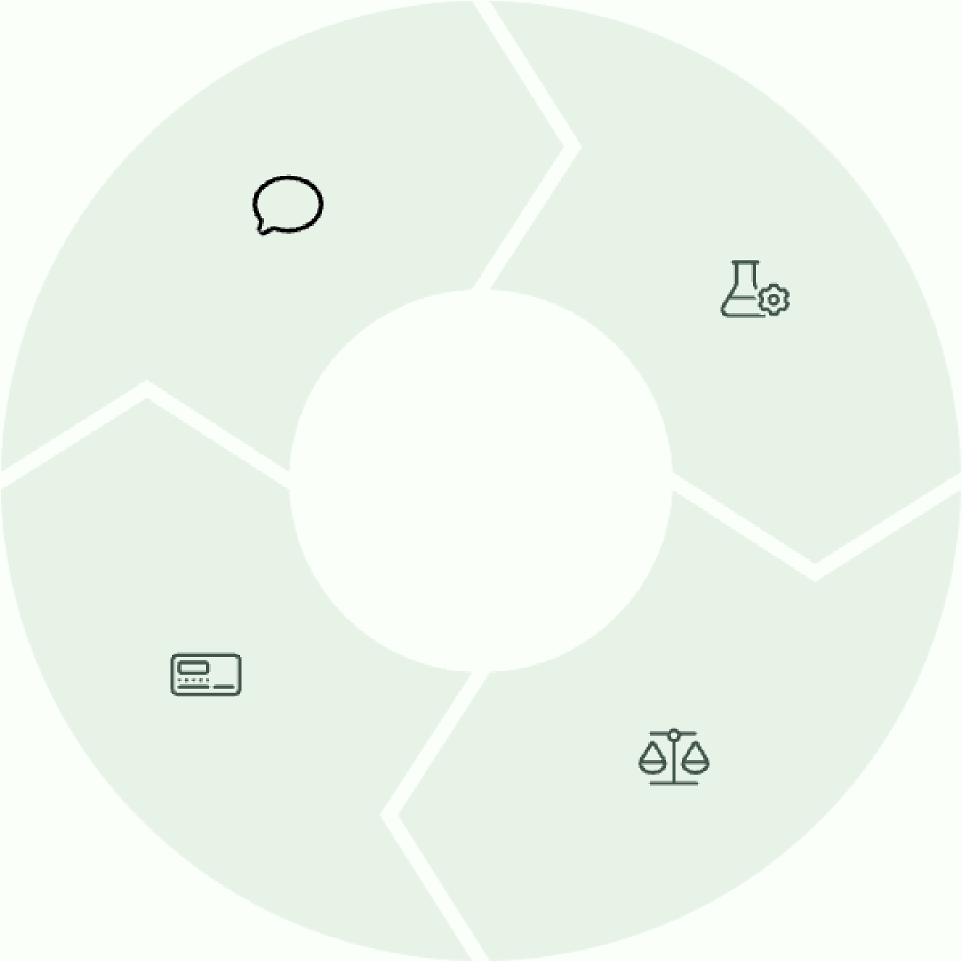




Clustering Model Specifications

Variable Standardization

Eleven standardized numeric variables used, including Age, MonthlyIncome, NumberOfTrips, and PitchSatisfactionScore



Configuration

MAXITER=100 and CONVERGE=0.02 to ensure model stability

Cluster Balance

Most groups had well-balanced sizes (1000+ records)

Separation

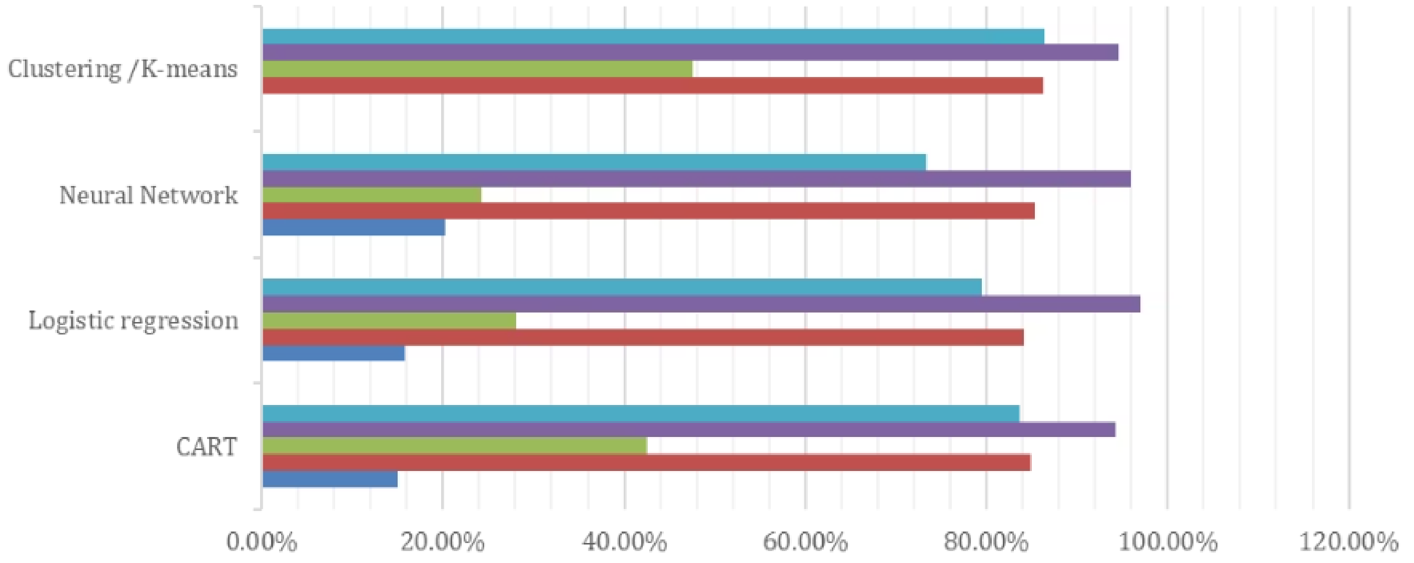
Centroid distances exceeded 1.98 across all clusters

The k=6 model achieved the highest pseudo-F-statistic (227.01), a more stable CCC (-36.182), and a higher average cluster separation (≈7.21). These results, combined with large, interpretable cluster sizes, made the six-cluster model superior for customer segmentation.

Results and Discussion

Conclusion

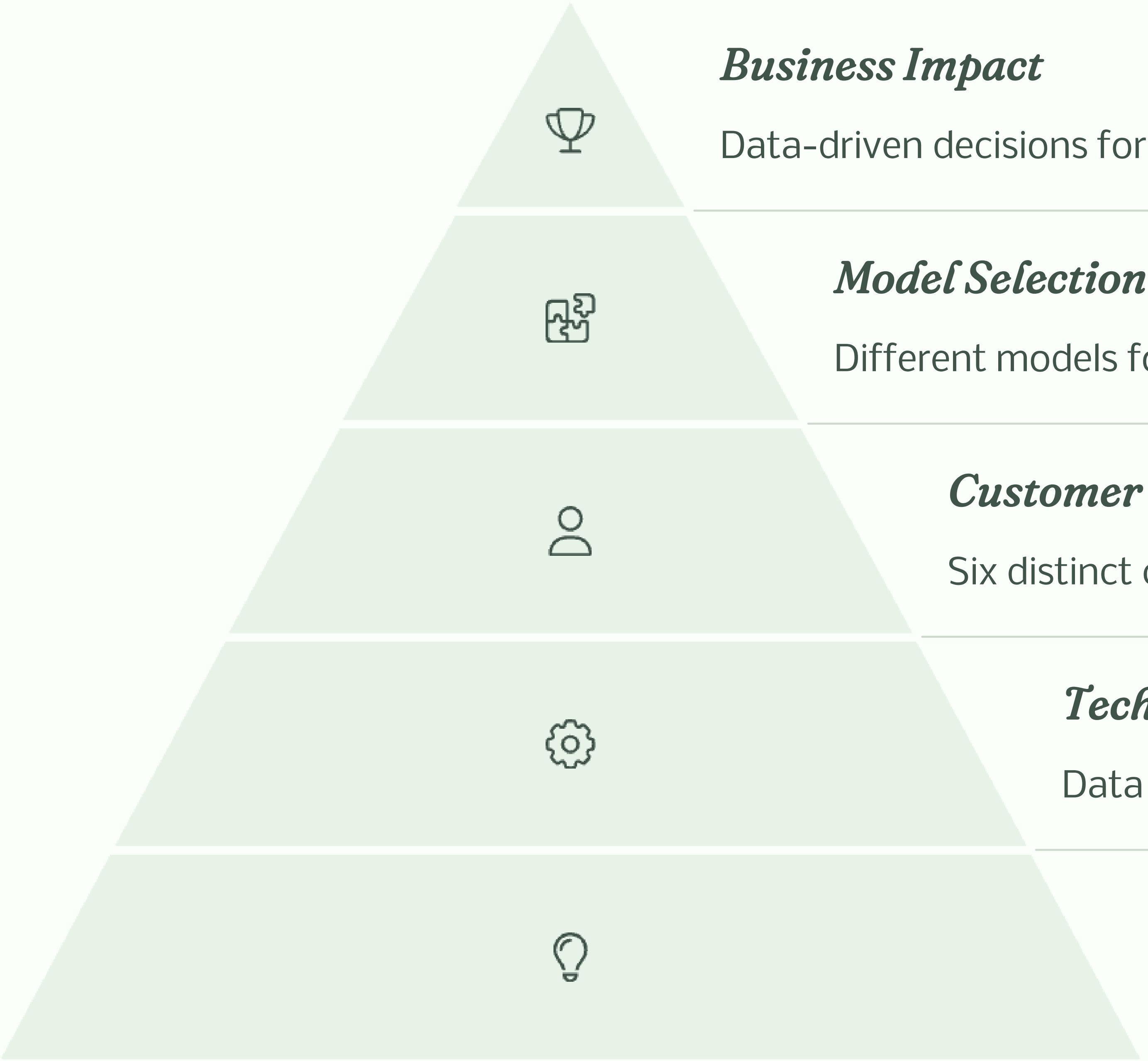
- CART with Entropy performed best overall
- K-means clustering added rich customer segmentation
- Logistic regression was stable but not sensitive enough
- Neural networks had mixed performance, useful for deep patterns



	CART	Logistic regression	Neural Network	Clustering /K-means
AUC	83.73%	79.60%	73.36%	86.50%
Specificity	94.28%	97.05%	96.00%	94.52%
Sensitivity	42.58%	28.02%	24.20%	47.65%
Accuracy	84.89%	84.19%	85.31%	86.27%
Misclassification	15.11%	15.81%	20.30%	0

Business Outcome: The models collectively enable data-driven decision making by improving conversion rates, enhancing campaign efficiency, and supporting personalized, cost-effective marketing strategies.

Summary and Key Takeaways



Business Impact

Data-driven decisions for targeted marketing

Model Selection

Different models for different business needs

Customer Segmentation

Six distinct customer groups identified

Technical Skills

Data cleaning, transformation, algorithm evaluation

Data Science Lifecycle

End-to-end implementation experience

Thank you!