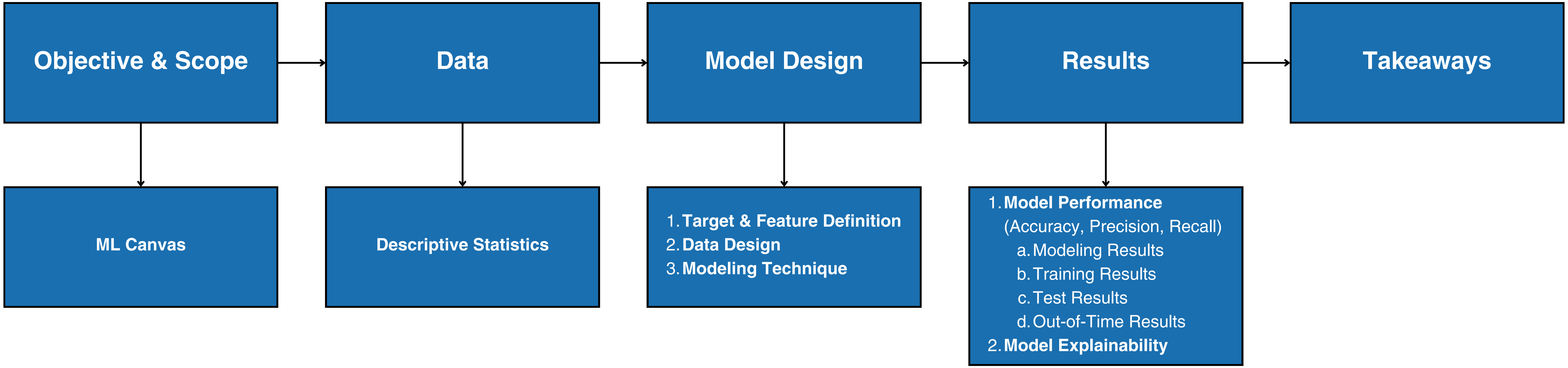
Machine Learning (ML) Project:

Predicting Customer Purchases to Increase Sales Revenue of a Low-Performing Product (Product 6850 - Gadgets)

PRESENTED BY
LUDREIN REIMAR R. SALVADOR

OVERVIEW


Objective & Scope
Data
Model Design
Results
Takeaways
ML Canvas
Descriptive Statistics
1. Target & Feature Definition
2. Data Design
3. Modeling Technique
1. Model Performance
(Accuracy, Precision, Recall)
a. Modeling Results
b. Training Results
c. Test Results
d. Out-of-Time Results
2. Model Explainability

OBJECTIVE AND HYPOTHESIS

Objective:

- To create a machine learning model that predicts whether a customer will purchase **Product 6850** in **January 2019** with an accuracy **above 85%** and a recall of at least **70%**.

Hypothesis:

- **Customer Demographics Impact Buying Behavior** - Customers of a certain age group, gender, or membership type are more likely to buy Product 6850.

- **Past Purchase History Predicts Future Purchases** - Customers who have purchased gadgets or similar product categories in the past 3 months are more likely to buy Product 6850.

- **Weather Conditions Affect Purchase Decisions** - Sales of Product 6850 are higher on cloudy/rainy days and lower on sunny days.

- **Frequent Shoppers are More Likely to Buy** - Customers with higher overall spending and transaction frequency have a higher chance of buying Product 6850.

OBJECTIVE & SCOPE | ML CANVAS

Important Links: [ML Canvas (Docs)] | [ML Canvas (PDF)]

OWNML MACHINE LEARNING CANVAS

Designed for: Machine Learning Designed by: Ludrein Reimar R. Salvador Date: 03/29/25 Iteration: 02

PREDICTION TASK

What is the type of task?
- Classification | Random Forest Classifier (Predict whether a customer will buy Product 6850 or not).

Which entity are predictions made on?
- Predictions are made on individual customers based on their past purchasing behavior.

What are the possible outcomes to predict?

- 1 (Yes) - The customer will purchase Product 6850.
- 0 (No) - The customer will not purchase Product 6850.

When are outcomes observed?
- **After each transaction period**, particularly in **January 2019**, based on customer behavior in the **previous three months (October–December 2018)**.

DECISIONS

How are predictions turned into actionable recommendations or decisions for the end-user? (Mention parameters of the process / application for this.)

- Predicted buyers can be targeted with personalized marketing campaigns (e.g., promotions, discounts).

- If a customer is predicted not likely to buy, additional strategies like bundling or cross-selling with other gadgets can be implemented.

Parameters:

- Customer demographics (age, gender, location)
- Past purchases (Gadget category spending, frequency of transactions)
- Weather conditions influencing sales behavior

VALUE PROPOSITION

Who is the end beneficiary, and what specific pain points are addressed?

- **Retail company's marketing team** → Can better **target potential buyers**, reducing wasted marketing efforts.

- **Retail customers** → Receive personalized product recommendations based on their interests.

- **Business management** → Increases **sales and revenue** for underperforming products like **Product 6850**.

How will the ML solution integrate with their workflow, and through which user interfaces?

- Predictions will be fed into the company's marketing platform to create targeted ad campaigns.

- Customer recommendations can be integrated into email marketing, app notifications, or personalized web banners.

DATA COLLECTION

How is the initial set of entities and outcomes sourced (e.g., database extracts, API pulls, manual labeling)?

- Data is sourced from the company's retail transaction database (**df_2018_demo_view** table in Spark SQL).

What strategies are in place to update data continuously while controlling cost and maintaining freshness?

- Automated data ingestion pipelines that pull new transactions every month.

- Periodic model retraining (every 3–6 months) to ensure predictive accuracy.

DATA SOURCES

Where can we get data on entities and observed outcomes? (Mention internal and external database tables or API methods.)

Internal sources:

- Retail transaction database (df_2018_demo_view)
- Customer demographics (age, gender, membership status)
- Purchase history (last 3 months)
- Product category spending
- Weather conditions on transaction days

External sources:

- Weather APIs (to get weather conditions per store location).
- Marketing data (response rates to previous promotions).

IMPACT SIMULATION

What are the cost/gain values for (in)correct decisions?

True Positive (TP): Predicting a customer will buy and they actually do → Increased revenue, better targeted marketing.

False Positive (FP): Predicting a customer will buy but they don't → Wasted marketing resources.

True Negative (TN): Predicting a customer will not buy and they don't → No unnecessary marketing spend.

False Negative (FN): Predicting a customer won't buy but they actually do → Lost revenue opportunity.

Which data is used to simulate pre-deployment impact?

Historical data from **October–December 2018** is used to train the model and test predictions for **January 2019**.

MAKING PREDICTIONS

Are predictions made in batch or in real time?

Predictions will be made in **batch mode** at the start of each month.

How frequently?

Every **3 months** (quarterly model retraining).

How much time is available for this (including featurization and decisions)?

1-2 hours per batch run, depending on data size and model complexity.

Which computational resources are used?

Google Colab / Spark clusters for training the model.

Retail company's data warehouse for deployment.

BUILDING MODELS

How many models are needed in production?

One main model (binary classification model for purchase prediction).

When should they be updated?

Every **3–6 months**, depending on how quickly customer behavior changes.

How much time is available for this (including featurization and analysis)?

1-2 days for retraining and validation.

Which computation resources are used?

Google Colab, Spark

FEATURES

What representations are used for entities at prediction time?

- Customer demographic data (age, gender, region, membership type).
- Spending behavior in different product categories (last 3 months).
- Transaction frequency (number of purchases per category).
- Weather conditions on past purchase dates.

What aggregations or transformations are applied to raw data sources?

- Sum, mean, min, max transactions per category.
- Past 3-month transaction history per customer.
- Categorical encoding (One-Hot Encoding for membership type, region).
- Scaling (Min-Max Normalization for numerical features).

OBJECTIVE & SCOPE | ML CANVAS

Important Links: ML Canvas (Docs) | ML Canvas (PDF)

What are the criteria for deployment?

Model must achieve **at least 85% accuracy** with a **high precision score (to avoid wasted marketing costs)**.

Are there fairness constraints?

The model should not **discriminate based on age, gender, or region**. Predictions should be purely based on transaction behavior.

MONITORING

Which metrics and KPIs are used to track the ML solution's impact once deployed, both for end-users and for the business?

- Accuracy, Precision, Recall
- Conversion rate (how many predicted buyers actually buy).
- Marketing ROI (cost per acquisition for targeted customers).
- Customer engagement with targeted promotions.

How often should they be reviewed?

- **Monthly**: Monitor **conversion rates** of predicted buyers.
- **Quarterly**: Review **model performance and retrain if needed.**

(Continuation)

DATA PRE-PROCESSING

product	transaction_count	total_sales
9980	103	669500
6850	103	288400
7236	105	745500
7335	105	273000
7451	106	763200
7213	107	866700
6840	109	991900
7405	112	873600
6791	113	248600
6836	113	858800

Low-Performing Products for 'GADGETS'

- Identified low-selling gadgets, focusing on Product 6850.
- Evaluated transaction frequency and revenue across gadget categories.
- From this, to objective is to find purchase patterns to improve sales.

DATA PRE-PROCESSING

age	gender	purchase_count
50-55	FEMALE	17
55-60	FEMALE	12
55-60	MALE	11
50-55	MALE	9
45-50	FEMALE	8
35-40	FEMALE	7
18-25	FEMALE	6
60 Above	FEMALE	6
40-45	FEMALE	6
30-35	FEMALE	4
25-30	MALE	3
45-50	MALE	3
25-30	FEMALE	3
35-40	MALE	3
30-35	MALE	2
40-45	MALE	2
60 Above	MALE	1

Customer Demographics

- Features: Considered age, gender, member status in purchase behavior.
- Certain age groups and genders showed higher gadget purchases.

DATA PRE-PROCESSING

branch_name	weather_list	card_type
G_Store	Partly sunny w sh...	REGULAR
F_Store	Mostly clear	REGULAR
F_Store	Partly cloudy	PREMIUM
D_Store	Partly sunny w sh...	PREMIUM
E_Store	Mostly cloudy w s...	REGULAR
G_Store	Cloudy	REGULAR
H_Store	Rain	REGULAR
G_Store	Showers	REGULAR
C_Store	Sunny	REGULAR
D_Store	Mostly clear	REGULAR
G_Store	Thunderstorms	REGULAR
H_Store	Sunny	REGULAR
A_Store	Partly cloudy	REGULAR
A_Store	Mostly sunny	PREMIUM
C_Store	Partly sunny	REGULAR
J_Store	Thunderstorms	REGULAR
G_Store	Clear	REGULAR
I_Store	Cloudy	REGULAR
I_Store	Sunny	PREMIUM
F_Store	Mostly cloudy w t...	PREMIUM

only showing top 20 rows

Features Variability

- Features Used: Branch, weather conditions, card type, spending amount.
- The goal of this features variability is to identify how external factors impact buying behavior.
- These outcomes were incorporated into feature engineering for ML model.

DATA PRE-PROCESSING

member_type	branch	age	gender	quantity	amount	card_type	weather_list	temperature_list	realfell_list
member	I	50-55	MALE	12	2800	REGULAR	Cloudy	28	32
member	D	50-55	MALE	18	2800	REGULAR	Partly sunny	29	33
member	A	35-40	MALE	9	2800	REGULAR	Mostly cloudy w t...	32	36
member	B	50-55	FEMALE	16	2800	REGULAR	Mostly cloudy w s...	33	37
member	E	50-55	FEMALE	16	2800	REGULAR	Cloudy	32	36

only showing top 5 rows

Preparing Data for ML Model

- Target Definition: **1** = Bought Product 6850, **0** = Bought other gadgets.
- Limited non-buyers to balance dataset.
- Encoded categorical variables, normalized numerical features.

DATA - DESCRIPTIVE STATISTICS

	quantity	amount	temperature_list	realfell_list	target
count	603.000000	603.000000	603.000000	603.000000	603.000000
mean	10.187396	4586.733002	26.295191	30.295191	0.170813
std	5.430905	2687.646732	5.247916	5.247916	0.376658
min	1.000000	100.000000	18.000000	22.000000	0.000000
25%	5.500000	2800.000000	22.000000	26.000000	0.000000
50%	10.000000	3700.000000	27.000000	31.000000	0.000000
75%	15.000000	6800.000000	31.000000	35.000000	0.000000
max	19.000000	9900.000000	35.000000	39.000000	1.000000

Take Aways from the Descriptive Statistics

- Identified patterns in spending behavior.
- Top Features: Age, quantity, amount spent, and weather affected purchases.
- This is to validate data before model training.

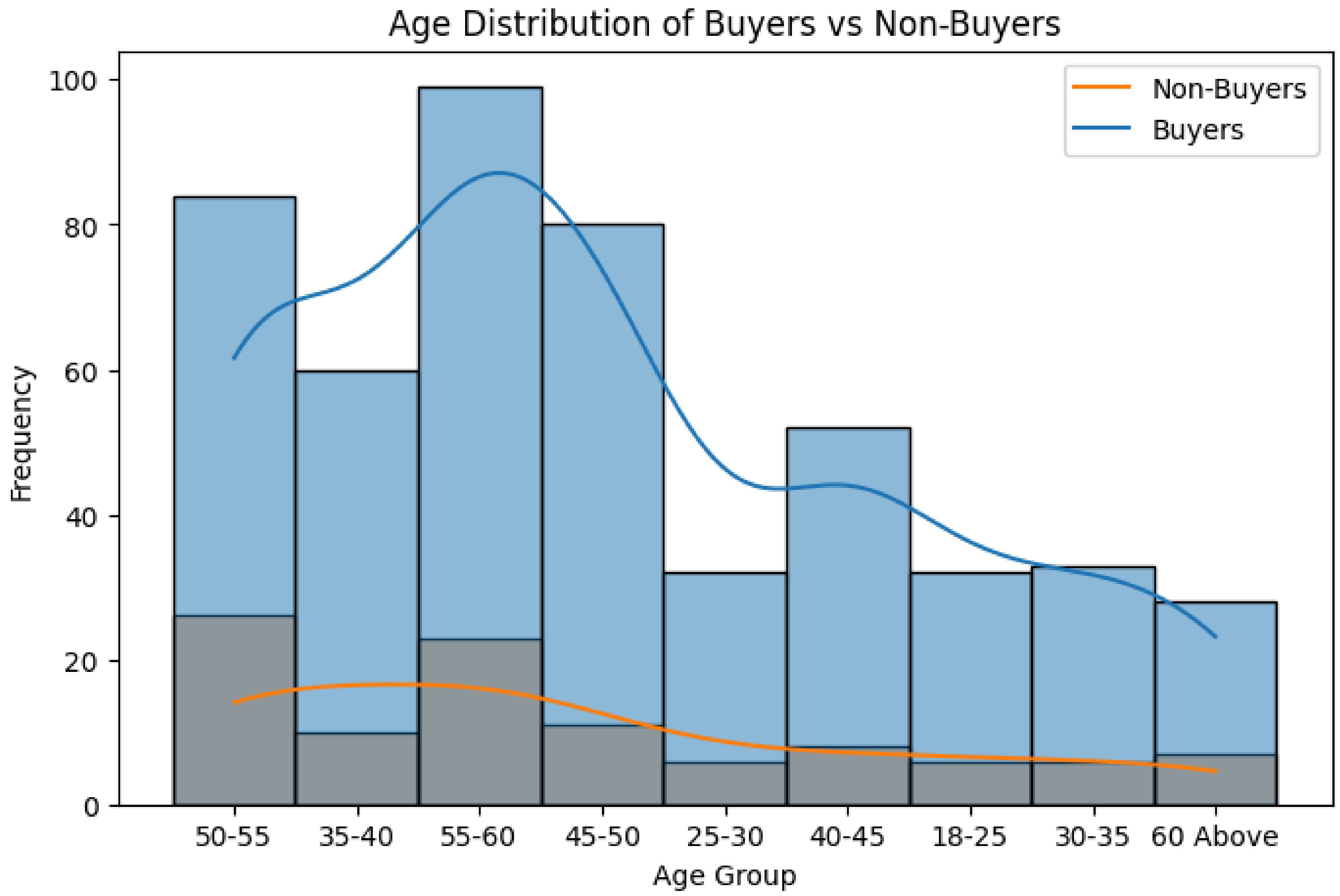
DATA - DESCRIPTIVE STATISTICS

```
(         quantity  amount  temperature_list  realfell_list  target
 count  103.000000   103.0        103.000000     103.000000   103.0
 mean     9.640777  2800.0         26.223301      30.223301     1.0
 std      5.586844     0.0          5.426753       5.426753     0.0
 min      1.000000  2800.0         18.000000      22.000000     1.0
 25%      5.000000  2800.0         22.000000      26.000000     1.0
 50%     10.000000  2800.0         27.000000      31.000000     1.0
 75%     15.000000  2800.0         31.000000      35.000000     1.0
 max     19.000000  2800.0         35.000000      39.000000     1.0,
        quantity       amount  temperature_list  realfell_list  target
 count  500.00000   500.000000        500.000000     500.000000   500.0
 mean    10.30000  4954.800000         26.310000      30.310000     0.0
 std      5.39706  2814.206189          5.215775       5.215775     0.0
 min      1.00000   100.000000         18.000000      22.000000     0.0
 25%      6.00000  2800.000000         22.000000      26.000000     0.0
 50%     11.00000  4600.000000         27.000000      31.000000     0.0
 75%     15.00000  7500.000000         31.000000      35.000000     0.0
 max     19.00000  9900.000000         35.000000      39.000000     0.0)
```

Take Aways from the Descriptive Statistics

- Identified patterns in spending behavior.
- Top Features: Age, quantity, amount spent, and weather affected purchases.
- This is to validate data before model training.

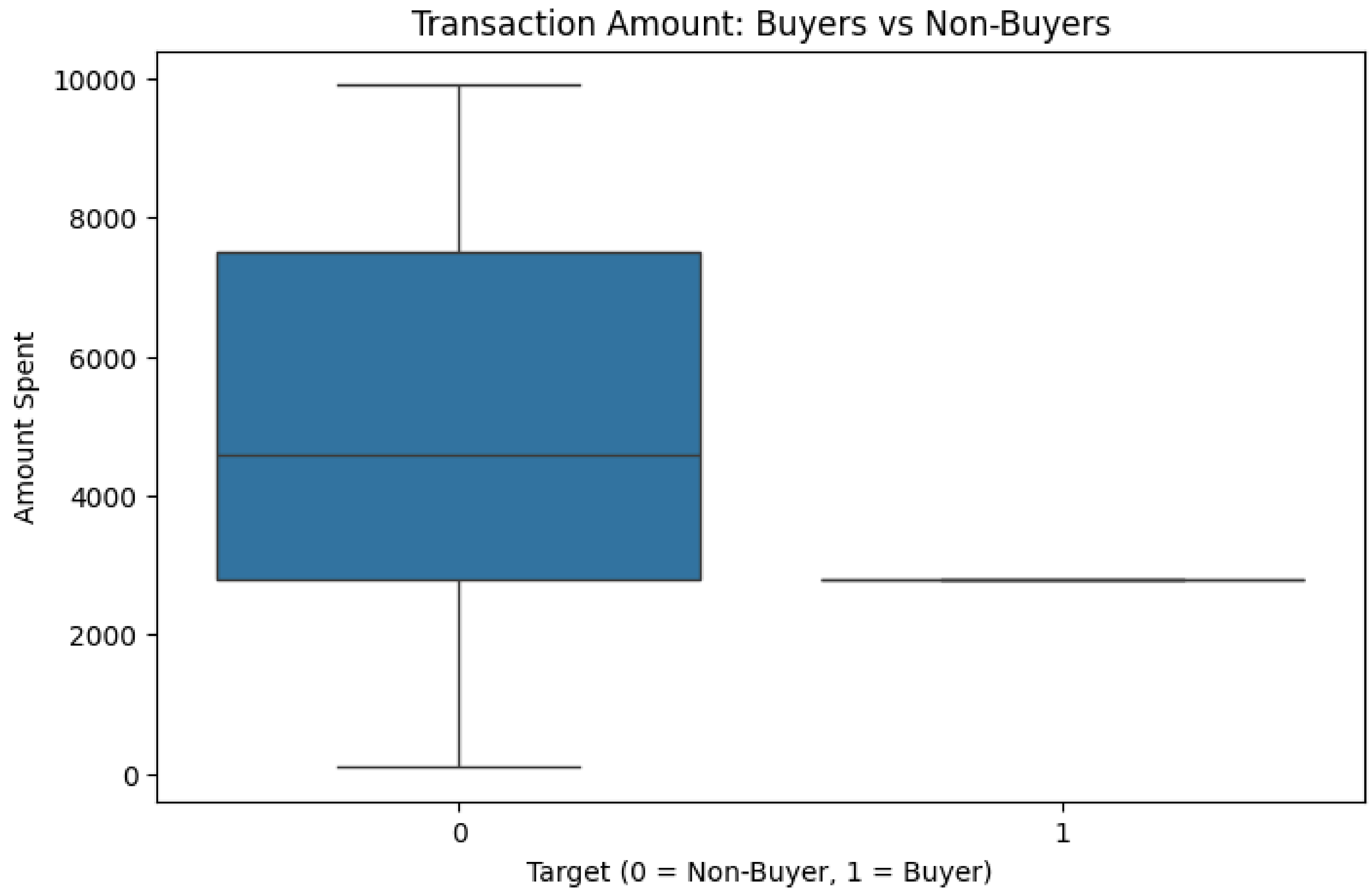
CHARTS - DESCRIPTIVE STATISTICS



Age Distribution of Buyers vs Non-Buyers

- Specific age groups dominate purchases.
- Older demographics (45-60 years) buy more gadgets.
- This would help refine marketing strategies.

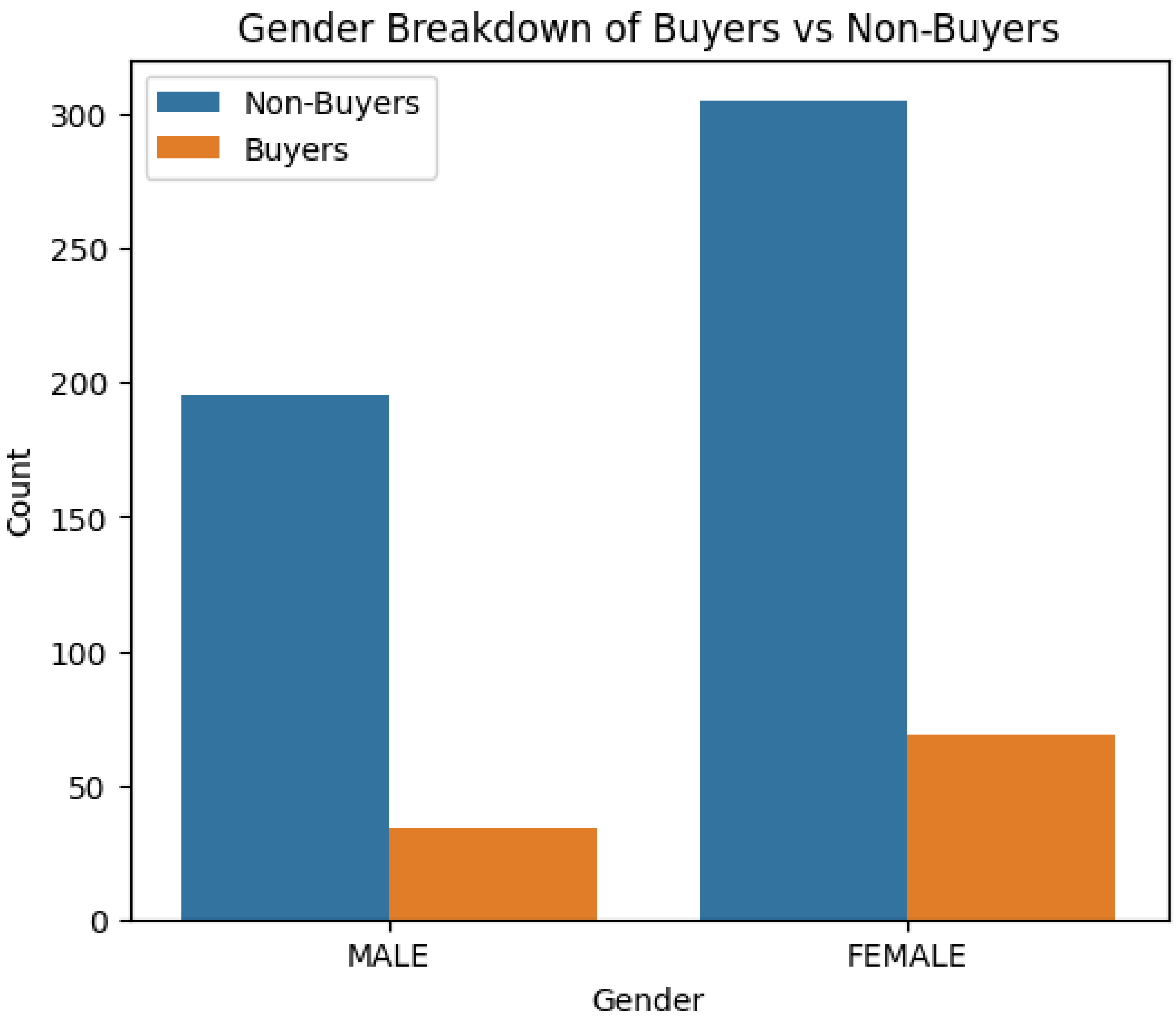
CHARTS - DESCRIPTIVE STATISTICS



Transaction Amount Comparison

- Buyers spend consistent amounts, indicating budget preference.
- Non-buyers have higher variance in spending.
- This helps in customer segmentation for targeted promotions.

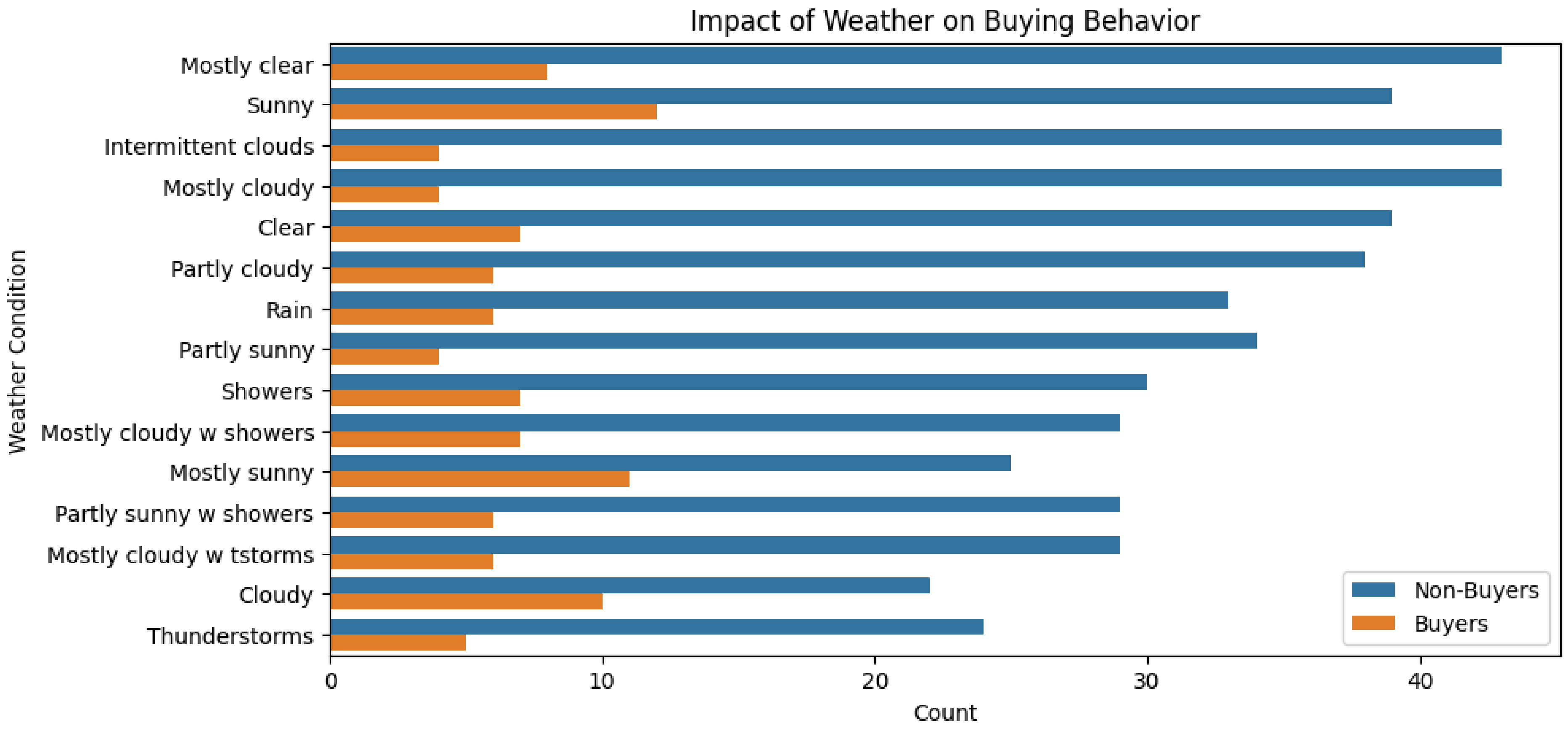
CHARTS - DESCRIPTIVE STATISTICS



Gender Comparison

- Female buyers slightly outnumber male buyers for Product 6850.
- Gender-based targeting may improve marketing effectiveness.
- This supports personalized recommendation strategies.

CHARTS - DESCRIPTIVE STATISTICS



Impact of Weather on Buying Behavior

- Trend: Cloudy/rainy days boost gadget sales.
- Hypothesis Validated: Weather affects consumer decisions.
- Seasonal promotions could improve sales.

MODEL DESIGN

Goal: Identify customers likely to purchase **Product 6850** (a gadget) based on **spending behavior over the past 3 months**.

2018			2019
3			1
OCT	NOV	DEC	JAN
Behavioral Window			Performance Window
Features			Target

Used to extract features that describe customer behavior.

Based on

- Total amount spent (all products)
- Amount spent on gadgets
- Avg, min, max transaction amount
- (sum_amount_gadgets, sum_amount_total, avg_amount_total, min_amount_total, max_amount_total
- Transaction count
- Customer Age & Gender
- Weather conditions during transactions.

This period is used to determine the target outcome (whether a customer bought Product 6850 or not).

Based on

- 1 → Customer **bought** Product 6850.
- 0 → Customer **did not buy** Product 6850.

MODEL

Modeling Technique

- Plan: Random Forest Classifier
  - Optimized for precision and recall.
  - Feature encoding, train-test split, and model training.

PREDICTON

EVALUATION

Assess the model’s accuracy, precision, and recall.

TRAINING THE MODEL

```
+------+----------+--------------------+
|target|prediction|         probability|
+------+----------+--------------------+
|     1|       1.0|[0.34218463777810...|
|     1|       1.0|[0.39314649434174...|
|     1|       1.0|[0.35386921502400...|
|     1|       1.0|[0.25629097368401...|
|     1|       1.0|[0.49269083988624...|
|     1|       1.0|[0.42562920455818...|
|     1|       1.0|[0.39645561748926...|
|     1|       1.0|[0.36578389857118...|
|     1|       1.0|[0.24260468245727...|
|     1|       1.0|[0.26488651400613...|
+------+----------+--------------------+
only showing top 10 rows
```

A. Modeling Results + B. Training Results

Creating Pipeline, Train-Test Splitting, and Training, and Predicting

- Created ML pipeline, processed data, and trained the model.
- Train-Test Split: 80% training, 20% testing.
- This is to achieve high recall for correct product predictions.

```
ROC-AUC Score: 0.992283950617284
+------+----------+-----+
|target|prediction|count|
+------+----------+-----+
|     1|       1.0|   18|
|     0|       0.0|  106|
|     0|       1.0|    2|
+------+----------+-----+
```

A. Test Results

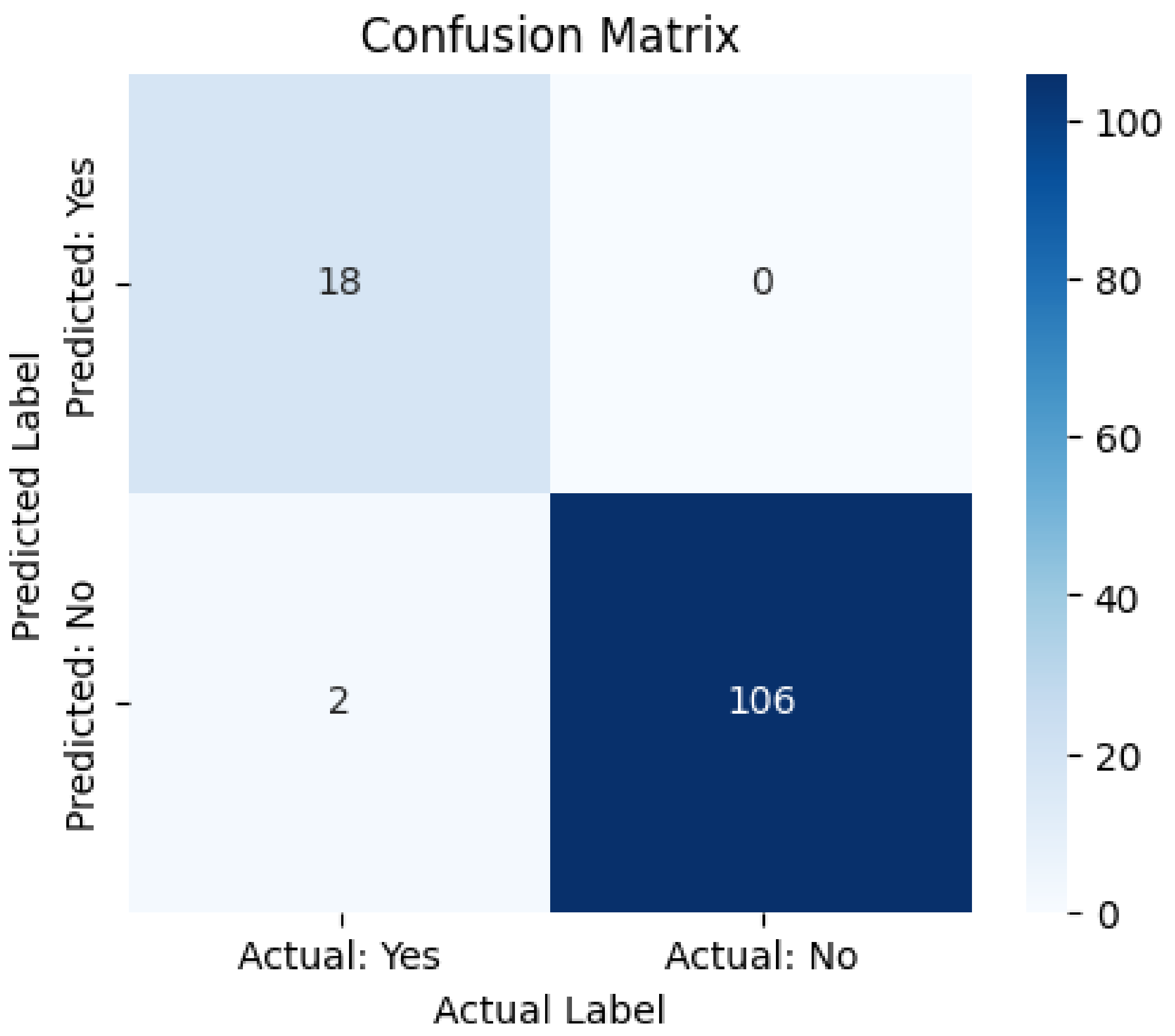
Evaluating the Model

- ROC-AUC Score: 0.99, this confirms strong classification ability.

- Demographics & spending history were top predictors.

RESULTS - MODEL PERFORMANCE

```
Accuracy: 0.9841
Precision: 0.9000
Recall: 1.0000
```



A. Test Results

Model Performance (Accuracy, Precision, Recall), Confusion Matrix

- Accuracy: 98.41%, Precision: 90%, Recall: 100%.
- Confusion Matrix: Low false negatives, high precision.
- Outcome: Model performs well in identifying potential buyers.

RESULTS - MODEL PERFORMANCE

```
20250329_0336
6850_prediction_20250329_0336_aggregated_3_months
```

	transaction_date	member_type	branch	age	gender	quantity	amount	card_type	weather_list	temperature_list	realfell_list
0	2020-01-01	member	I	35-40	MALE	18	4600	REGULAR	Cloudy	25	27
1	2020-01-01	member	B	55-60	FEMALE	13	100	PREMIUM	Sunny	32	30
2	2020-01-02	member	J	60 Above	MALE	14	100	REGULAR	Rain	22	22
3	2020-01-03	member	G	45-50	FEMALE	6	2800	PREMIUM	Thunderstorms	30	27
4	2020-01-03	member	I	18-25	FEMALE	8	3700	REGULAR	Partly cloudy	27	24

```
+------------------------------------------------+
|features                                        |
+------------------------------------------------+
|[0.0,1.0,4.0,1.0,18.0,4600.0,0.0,5.0,25.0,27.0]|
|[0.0,0.0,8.0,0.0,13.0,100.0,1.0,0.0,32.0,30.0] |
|[0.0,7.0,9.0,1.0,14.0,100.0,0.0,1.0,22.0,22.0] |
|[0.0,9.0,6.0,0.0,6.0,2800.0,1.0,3.0,30.0,27.0] |
|[0.0,1.0,1.0,0.0,8.0,3700.0,0.0,2.0,27.0,24.0] |
+------------------------------------------------+
only showing top 5 rows
```

D. Out-of-Time Results

Out-of-Time Results

- Validation: Applied model to 2020 transactions to test generalization.
- Performance remained consistent with training data.
- Continuous model retraining with new data.

TAKEAWAYS

- The model achieved high accuracy, making it effective for customer targeting.
- Business Impact - Helps improve sales by predicting high-potential buyers.

- **Future Improvements**
  - Deploy the model in real-time recommendations.
  - Monitor predictions and retrain with new data for continuous improvement.
  - Use insights for personalized marketing strategies.

Thank You!

PRESENTED BY
LUDREIN REIMAR R. SALVADOR