SUPERMARKET PROFIT PREDICTION PROJECT

PRESENTED BY:

ALYSHA PENMAN (T00740356)
APARNA RAJASEKAR (T00709748)
VIMAL ASTALAKSHMI PALANIAPPAN (T00717797)
KHUSHBOO CHRISTIAN (T00723895)

INTRODUCTION

This presentation covers a machine learning project aimed at predicting profit per transaction in supermarkets, utilizing the 'Supermarket Sales' dataset from Kaggle. The analysis applies Multiple Linear Regression and Decision Tree Regression models to derive actionable business insights.

PRIMARY GOALS

- Identify the key factors that influence transaction-level profit
- Build a machine learning model to predict profit accurately
- Compare the performance of Multiple Linear Regression and Decision Tree Regression
- Provide actionable business insights to support decisions in marketing, inventory, and operations
- Demonstrate how machine learning can be used for strategic value creation in the retail sector

DATA UNDERSTANDING

- dataset: The Supermarket Sales

- Size of the data is 1,000 transactions from 3 cities from (Jan–Mar 2019)

Key variables:
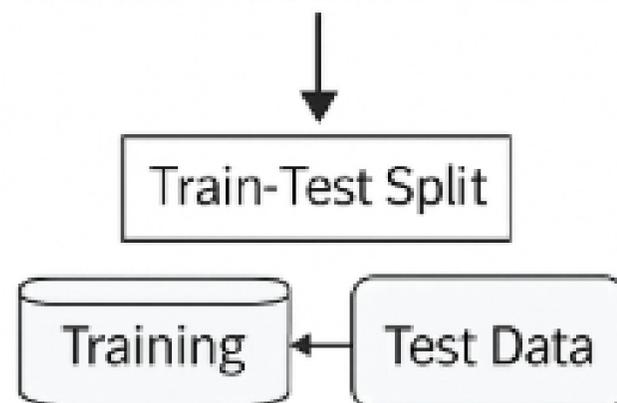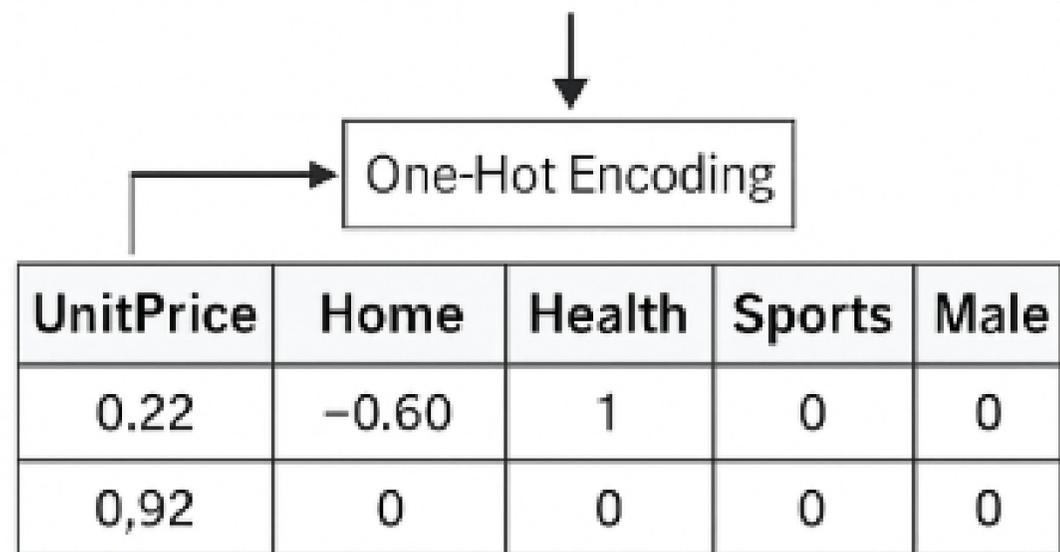- Unit Price, Quantity, Product Line, Customer Type, Payment Method, etc.

Target variable: (GOAL)
- Predicting Profit per transaction

DATA PREPARATION

City	ProductLine	Gender	UnitPrice		
Yangon	Home	Female	0	0	0
Mandalay	Health	Male	1	0	1
Yangon	Sports	Male	0	0	0

↓

One-Hot Encoding

UnitPrice	Home	Health	Sports	Male
0.22	−0.60	1	0	0
0,92	0	0	0	0

↓

Train-Test Split

↓

Training ← Test Data

Categorical Encoding:
- Converted features like Product Line, Gender, Customer Type into numbers using One-Hot Encoding(dummy coding)

Numerical Scaling:
- Applied Standard Scaler to normalize Unit Price and Quantity
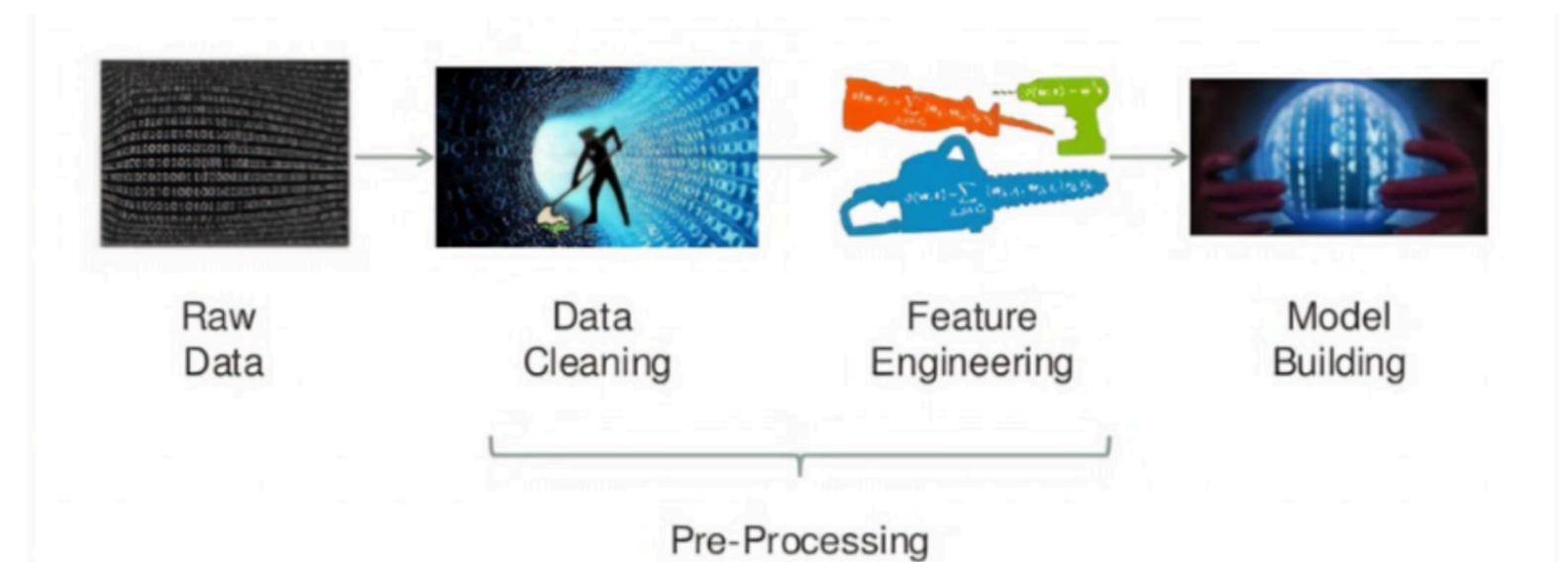
Datetime Formatting:
- Cleaned and converted Date and Time columns for consistency

Train-Test Split:
- Split dataset into 80% training and 20% testing

It ensures balanced and unbiased model evaluation

DATA CLEANING & FEATURE ENGINEERING

Scrub for Duplicate

Scrub for Irrelevant Data

Scrub for Incorrect Data

Fix Structural Errors

Handle Missing Data

Check the Outliers

Standardize

Normalize

Raw Data

Data Cleaning

Feature Engineering

Model Building

Pre-Processing

By that, we will be getting a clean, accurate, and analysis-ready dataset.

Exploratory Analysis

- Created a histogram of profit showing a right-skewed distribution

- Used a boxplot to identify outliers in profit values

- Built a correlation matrix revealing strong links between Total, COGS, and Profit

- Found that customer type, product line, and payment method influence profit outcomes

Histogram of Profit

- Distribution of transaction-level profit
- Right-skewed: most transactions generate $5–15 in profit
- Fewer transactions generate high profit (e.g., $30+)
- Indicates opportunity to increase profit per transaction


Histogram of Profit
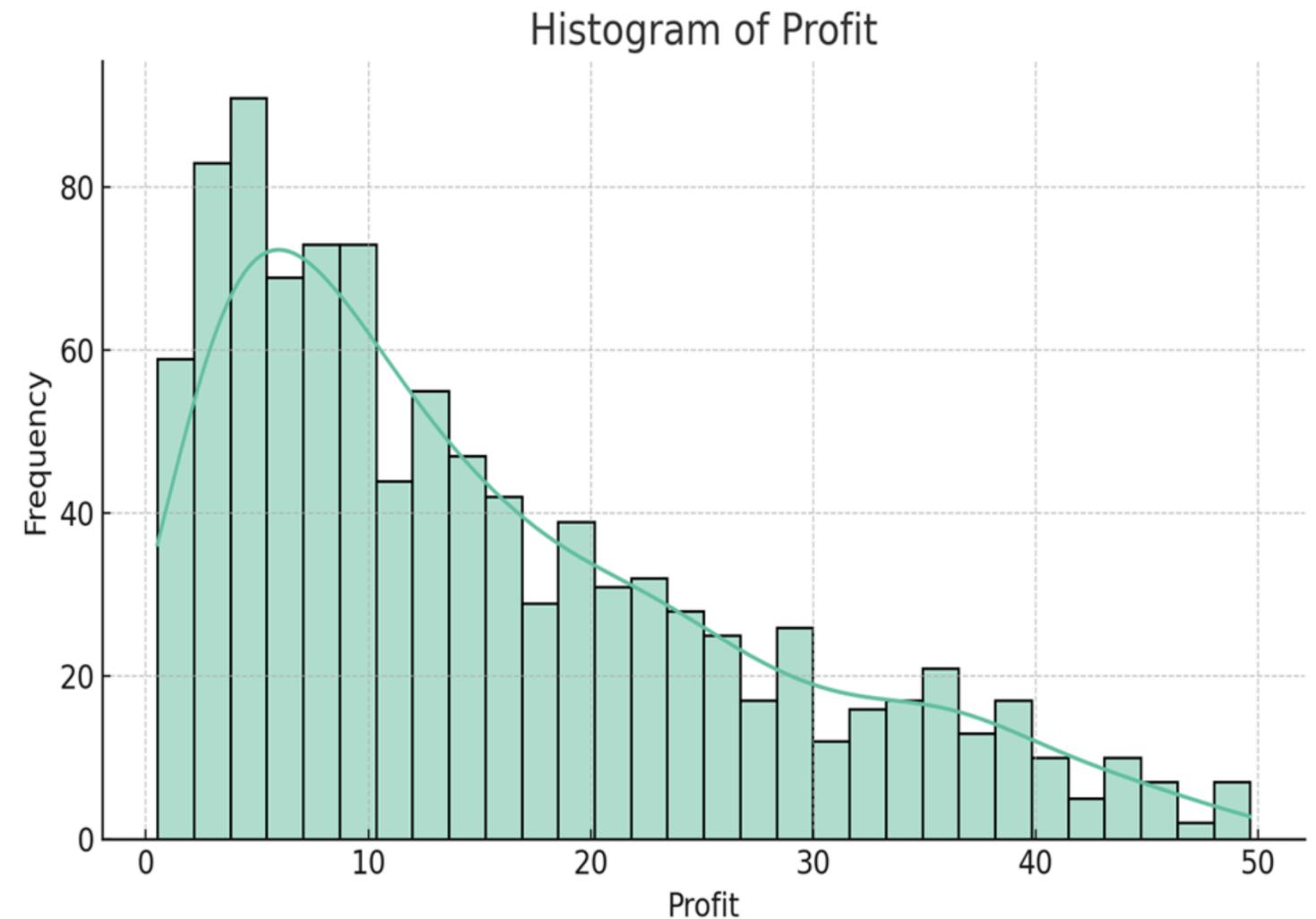
MODEL APPROACH

- Median profit is around $15 per transaction
- Most profits lie between $5 and $25
- A few high-profit outliers exist beyond $40
- Indicates variation in profitability per transaction

Boxplot of Profit
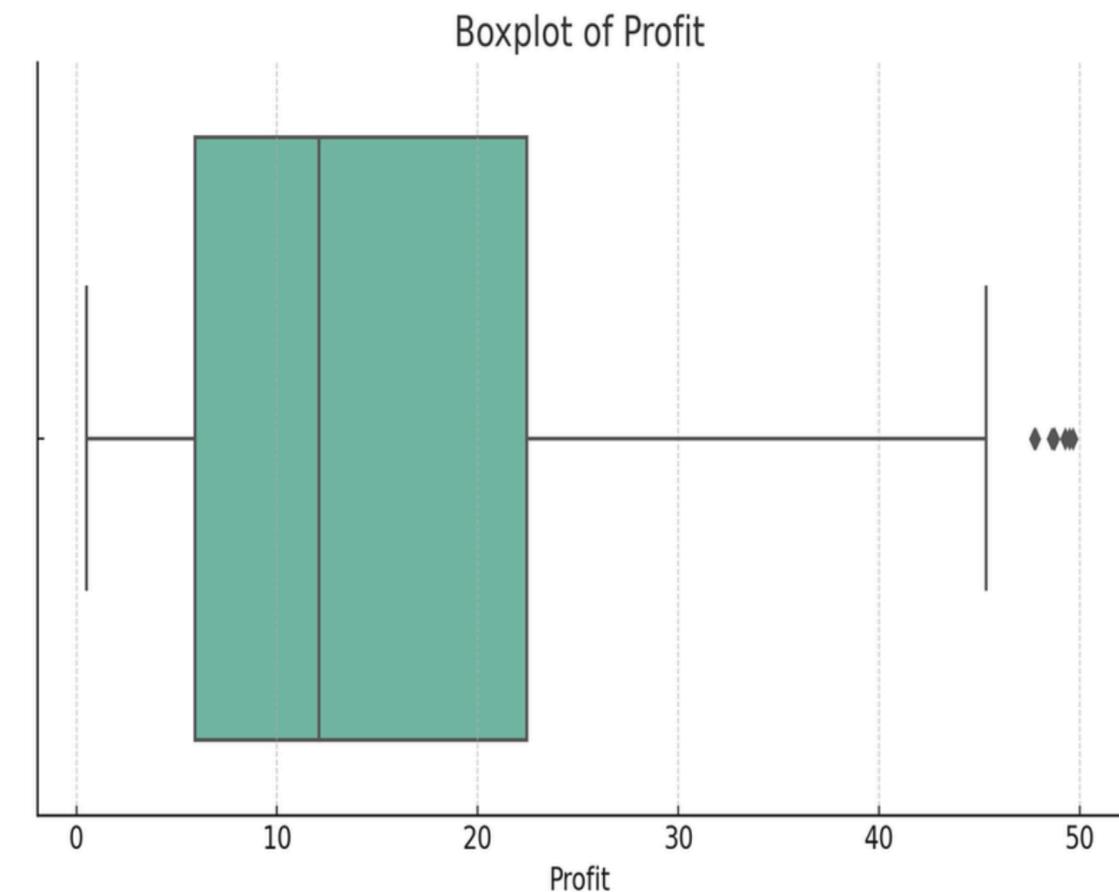
Profit

Actual vs Predicted Profit (Decision Tree)

- Compares predicted profit values vs actual profits from the test set
- Points close to red line = accurate predictions
- The red line represents perfect prediction (y = x)
- Model predictions closely follow this line → high accuracy
- Confirms the strong performance of Decision Tree Regression

Actual vs Predicted Profit (Decision Tree Regression)
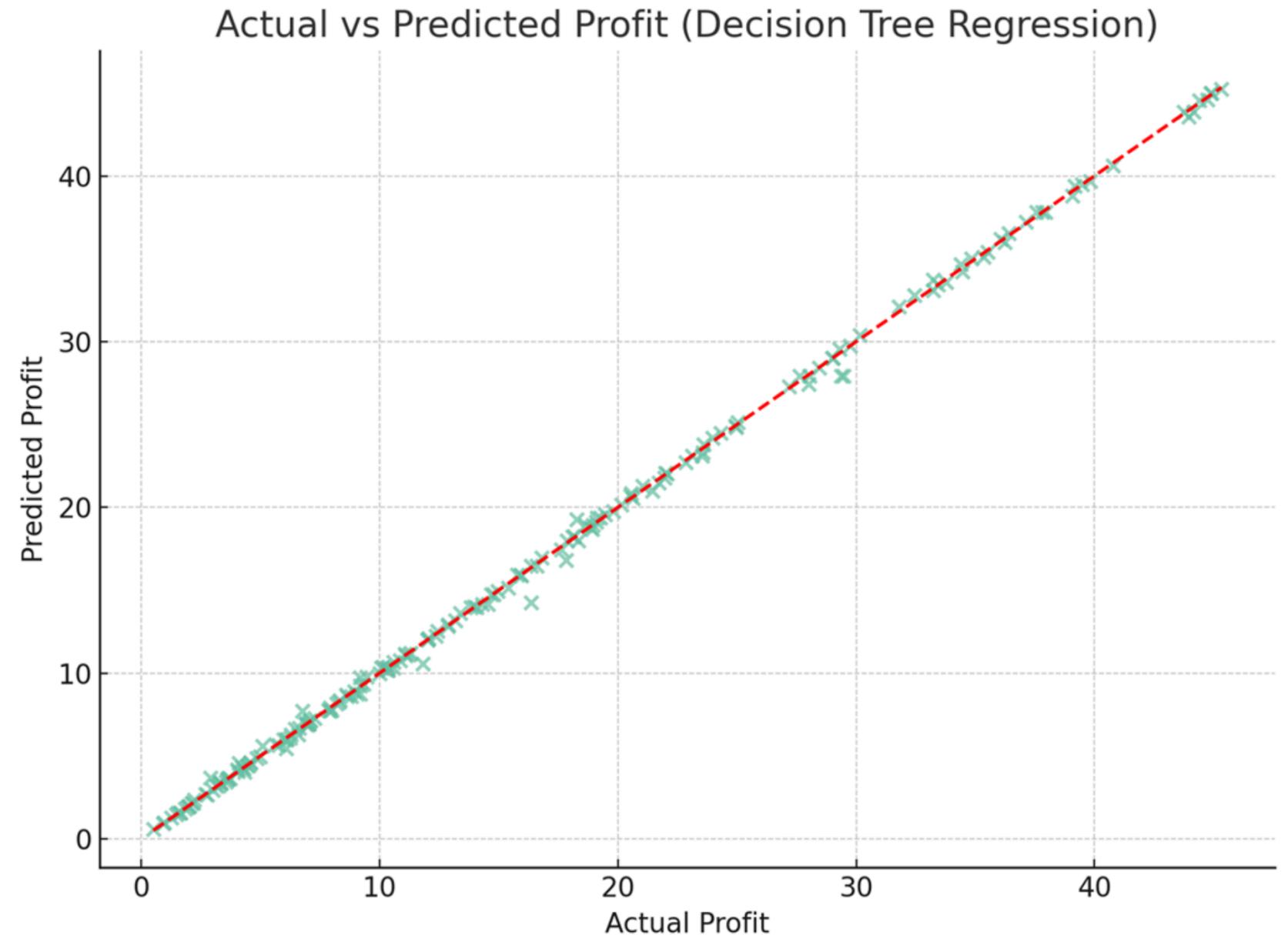
Final Result & Model Selection
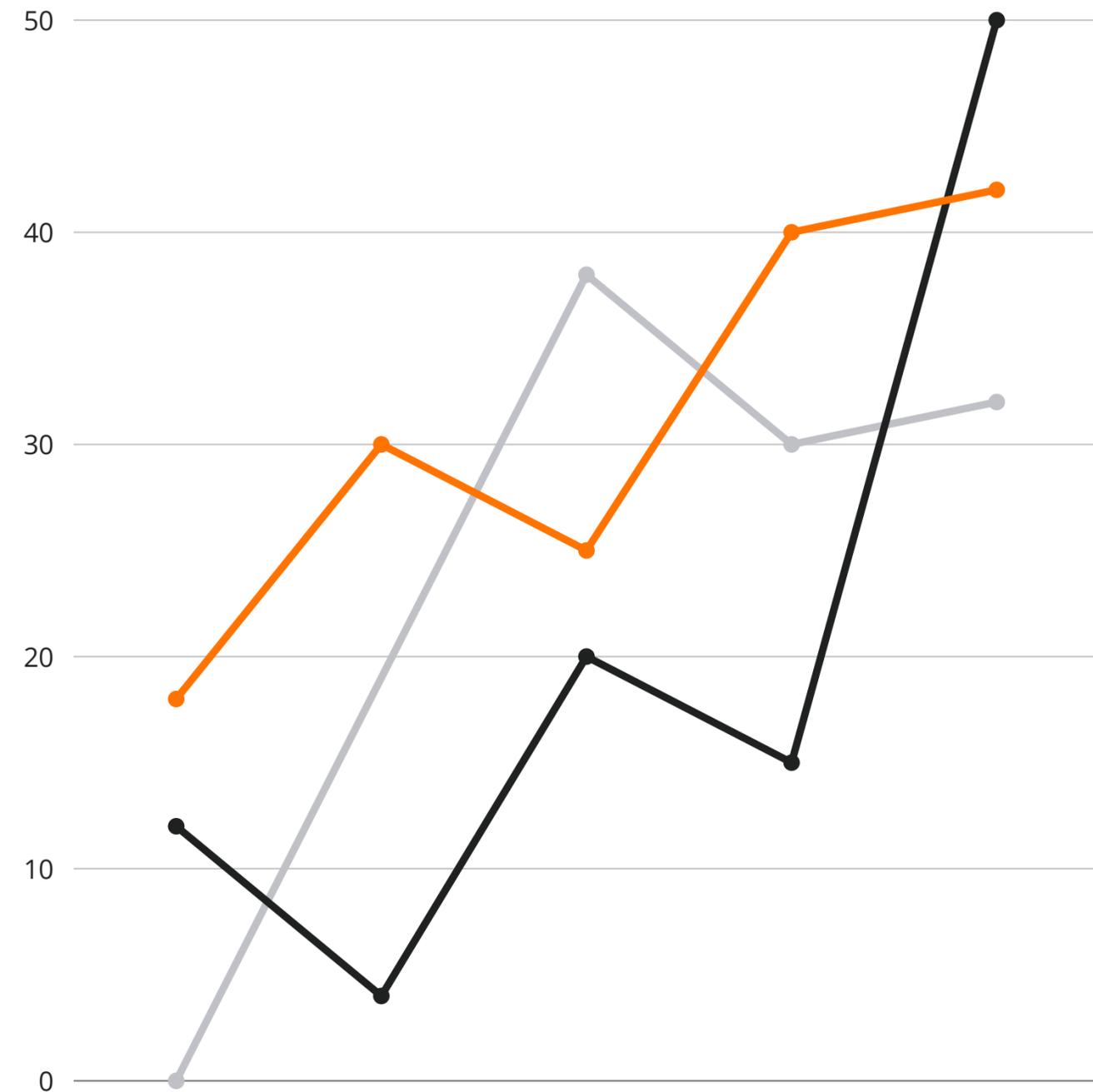
Multiple Linear Regression

R²: 0.902, RMSE: 3.79, MAE: 2.82. Captured linear relationships but less effective with complex patterns.
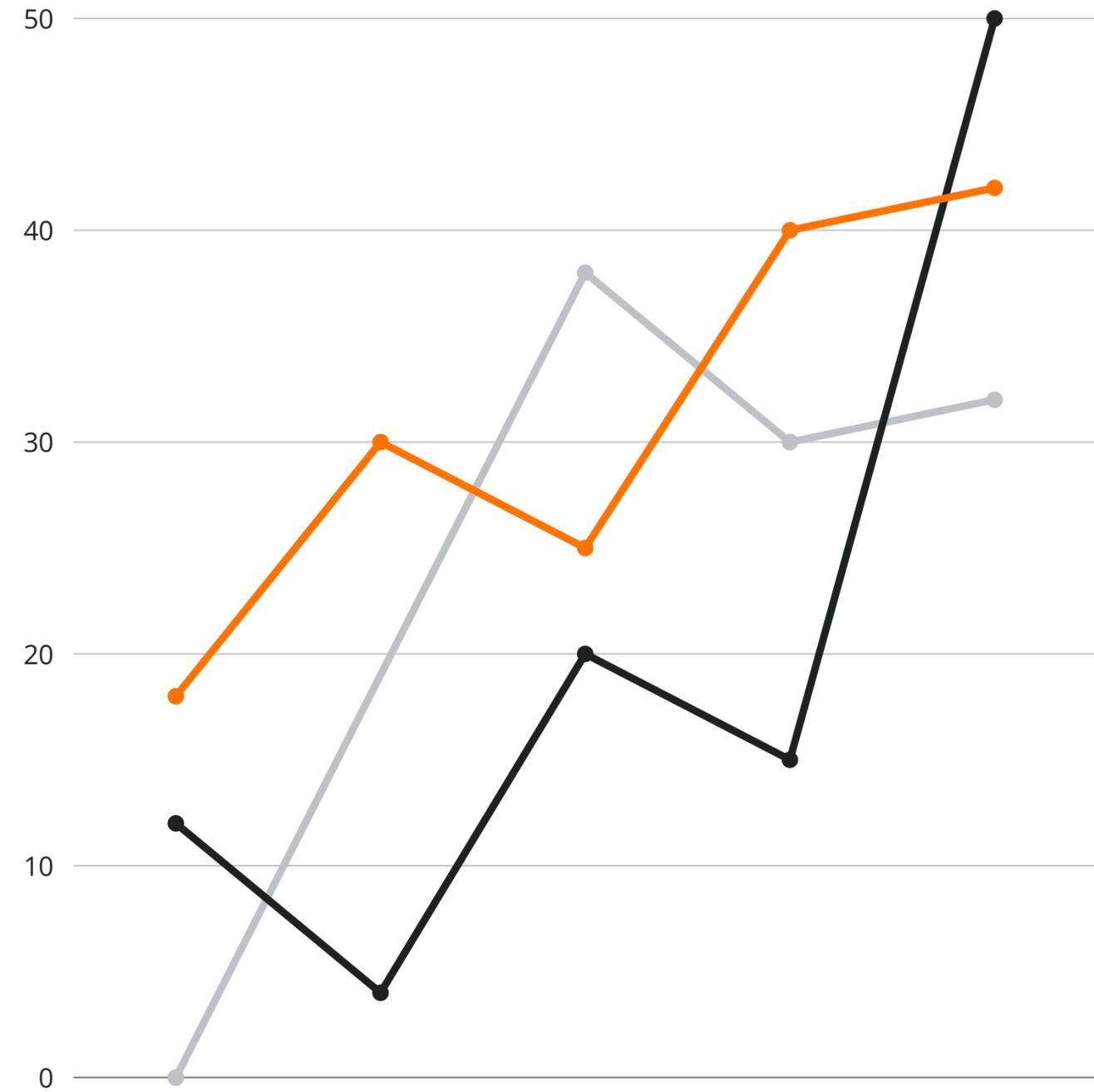
Decision Tree Regression

R²: 0.997, RMSE: 0.69, MAE: 0.43. Superior performance with high predictive accuracy and feature importance insights.
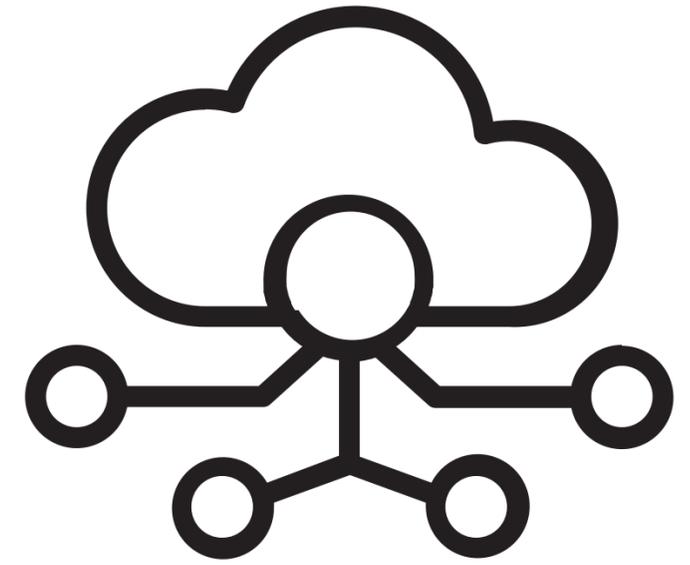
Final Result

Our Decision Tree model predicts profit with 99.7% accuracy and only $0.43 average error. That makes it a powerful tool for helping managers make smarter decisions around pricing, inventory, and sales strategy.
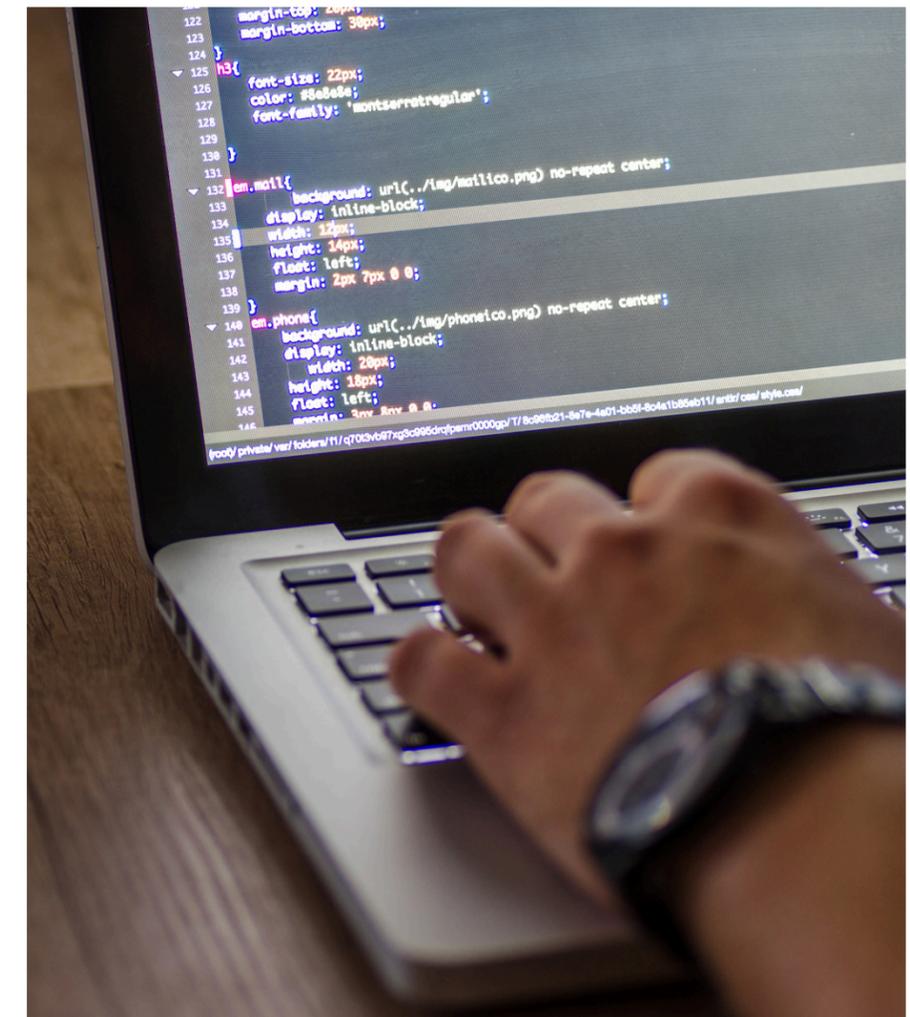
BUSINESS IMPLICATIONS

Operations:
- Allocate resources and staff during peak high-profit times
- Use predictions to optimize store hours and staffing schedules

Finance and Forecasting:
- Use profit prediction models to create more accurate revenue forecasts
- Plan budgets more effectively based on data-driven profit insights

Customer Retention:
- Reinforce loyalty programs for Member customers
- Use satisfaction and rating data to identify and retain high-value customers

BUSINESS IMPLICATIONS

Marketing:
- Focus campaigns on high-profit product lines like Home and Lifestyle and Food and Beverages
- Target Member customers and E-wallet users, who tend to generate higher profit per transaction

Sales Strategy:
- Train staff to upsell high-profit product combinations
- Encourage customer loyalty programs to increase return visits from profitable customer segments

Inventory Management:
- Stock more products from high-performing categories
- Adjust reorder levels based on predicted profit trends
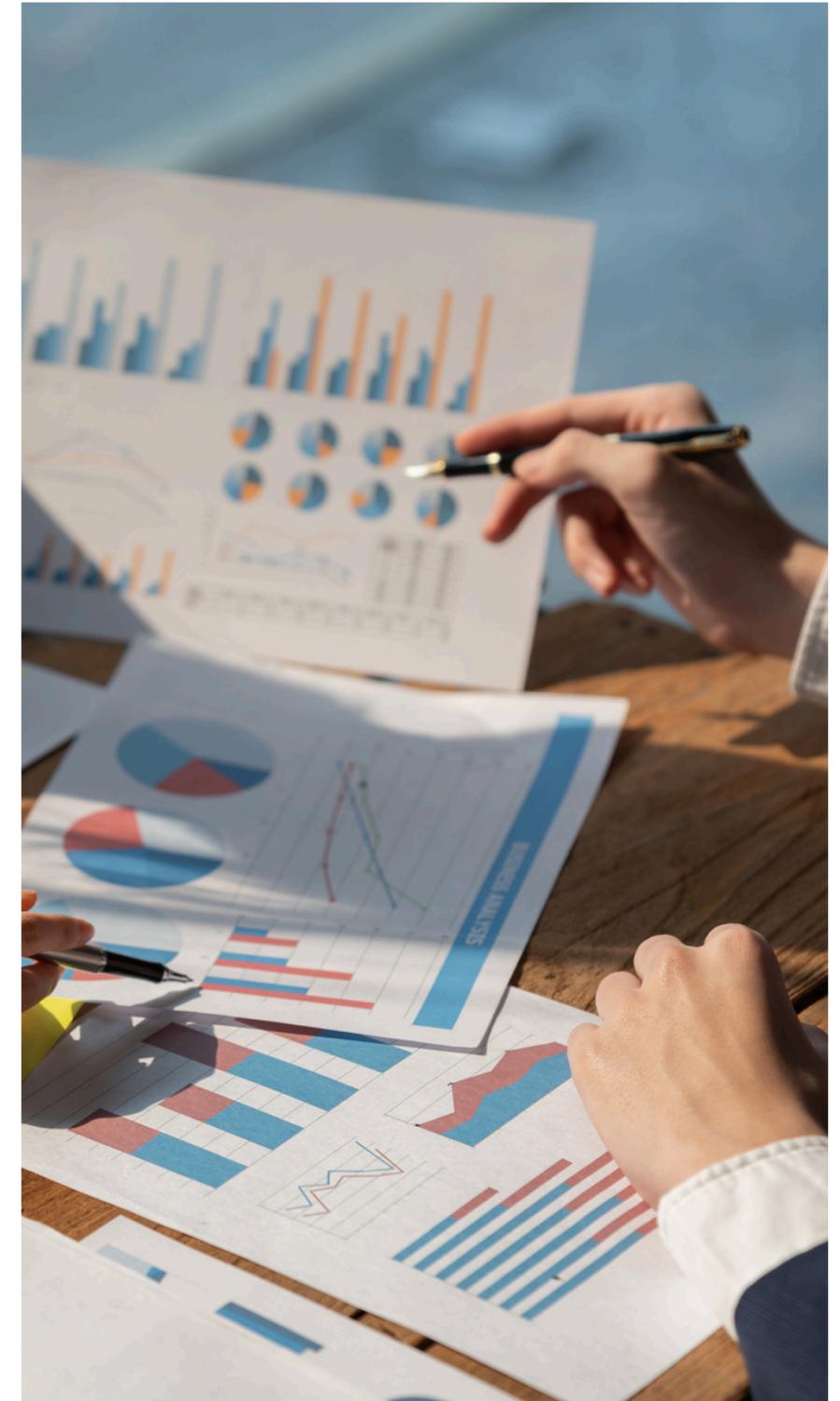
CONCLUSION

- Successfully built two predictive models to forecast transaction-level profit

- Decision Tree Regression outperformed Multiple Linear Regression with an R^2 of 0.997

- Key drivers of profit include unit price, quantity, product line, and customer type

- Cleaned and preprocessed data enabled accurate and reliable model performance

- Insights gained can directly support data-driven decision-making across departments

- Demonstrates the value of machine learning in retail strategy and planning

THANK YOU

ANY QUESTIONS