Data Analyst Project
2025

Quantity Sold E-Commerce Products Prediction

By: Muhammad Fakhri Azhar

Data Analyst Project
2025

Introduction

Hi! I'm Muhammad Fakhri Azhar, a physics graduate with a strong passion for data analysis. This project is part of my learning journey in turning data into insights.

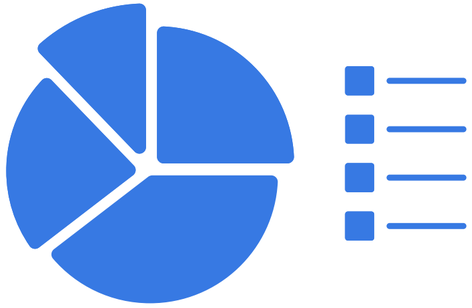
Course License :

- Data Science Bootcamp @Kelas Work by Kelas.com
- Data Analyst Mini Course @RevoU
- Ms.Excel Short Class @MySkill
- Computer Training @FMIPA UNNES

Contact Info

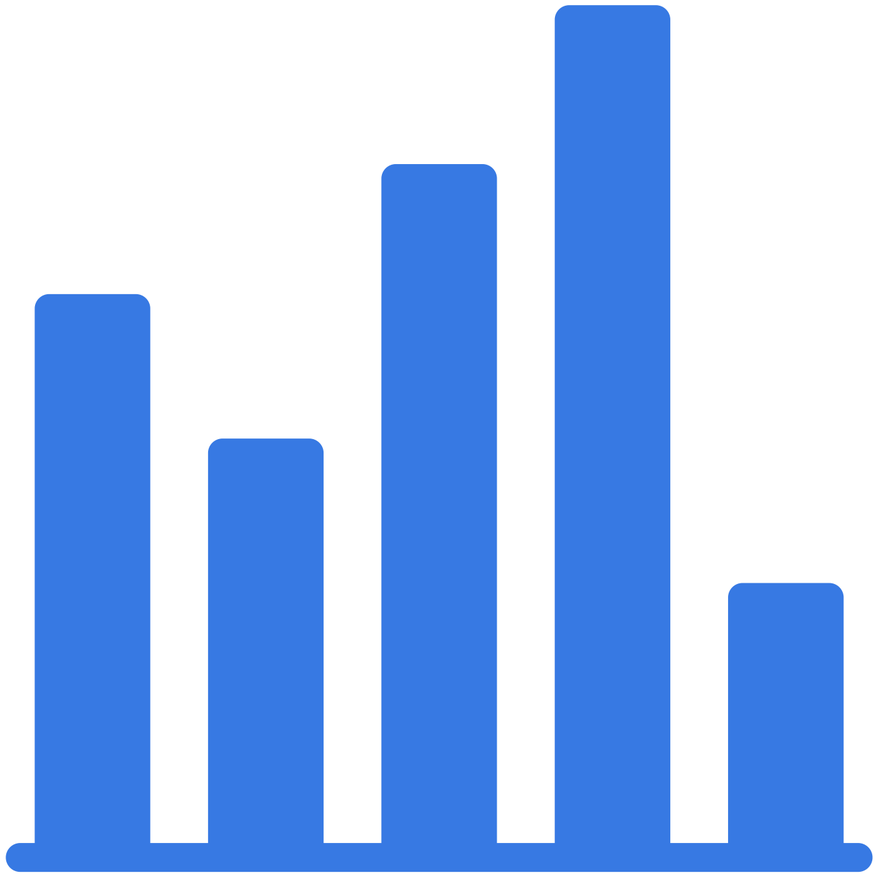
Email : mfkriazh57@gmail.com
Phone : 0857-2454-9367
LinkedIn : Muhammad Fakhri Azhar
Portfolio : Click here
GitHub : mfakhriazhar

Project Code Details on Github :
https://github.com/mfakhriazhar/ecom-qtt-prediction/blob/main/Case_03_Final_Project.ipynb



Overview

In e-commerce, understanding seasonal sales trends and best-selling products is critical to business strategy. However, companies often struggle with predicting sales, determining factors that influence sales (discounts, product categories, locations), and optimizing stock and marketing.



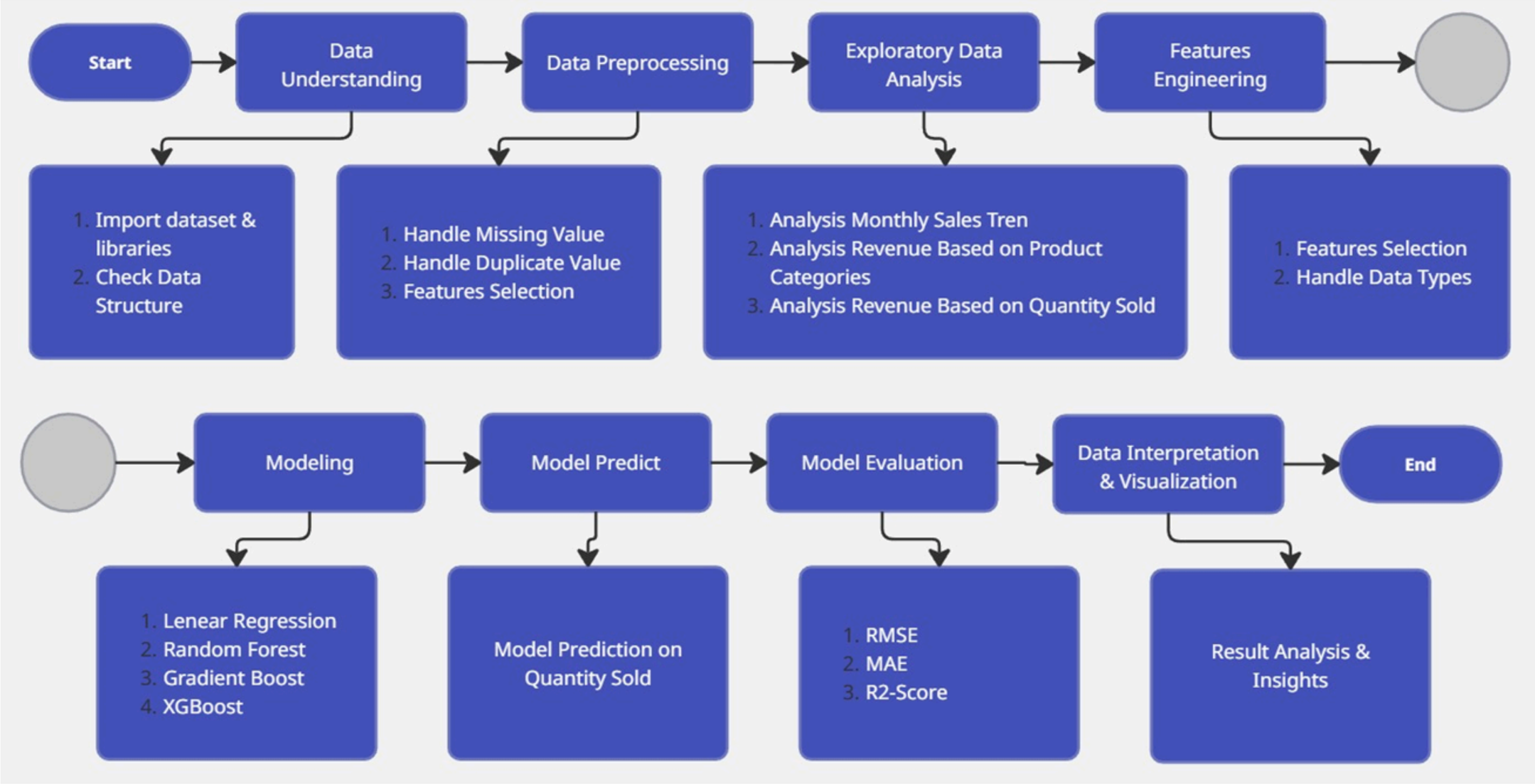
Dataset Link:

https://github.com/mfakhriazhar/ecom-qtt-prediction/blob/main/e%20commerce%20dataset.csv

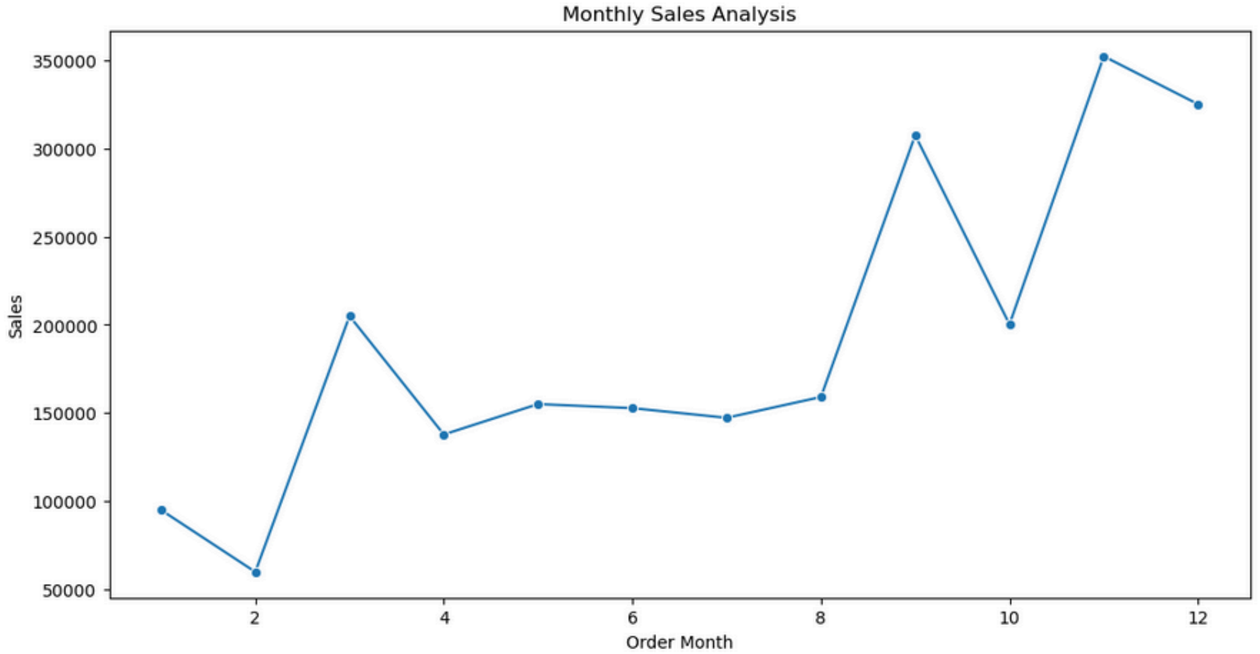
Objectives & Metodology

- Complex sales data requires in-depth analysis to identify buying patterns, determine high-demand products, and devise more effective pricing strategies. Therefore, this project aims to analyze sales trends, identify best-selling products, and build accurate prediction models.
- The methods used include data exploration and preprocessing, feature engineering, and application of ML models such as Linear Regression, Random Forest, Gradient Boosting, and XGBoost. The results are visualized to provide insights that can help optimize data-based business strategies.

Flowchart


Start
Data Understanding
Data Preprocessing
Exploratory Data Analysis
Features Engineering
1. Import dataset & libraries
2. Check Data Structure
1. Handle Missing Value
2. Handle Duplicate Value
3. Features Selection
1. Analysis Monthly Sales Tren
2. Analysis Revenue Based on Product Categories
3. Analysis Revenue Based on Quantity Sold
1. Features Selection
2. Handle Data Types
Modeling
Model Predict
Model Evaluation
Data Interpretation & Visualization
End
1. Lenear Regression
2. Random Forest
3. Gradient Boost
4. XGBoost
Model Prediction on Quantity Sold
1. RMSE
2. MAE
3. R2-Score
Result Analysis & Insights

Monthly Sales Analysis

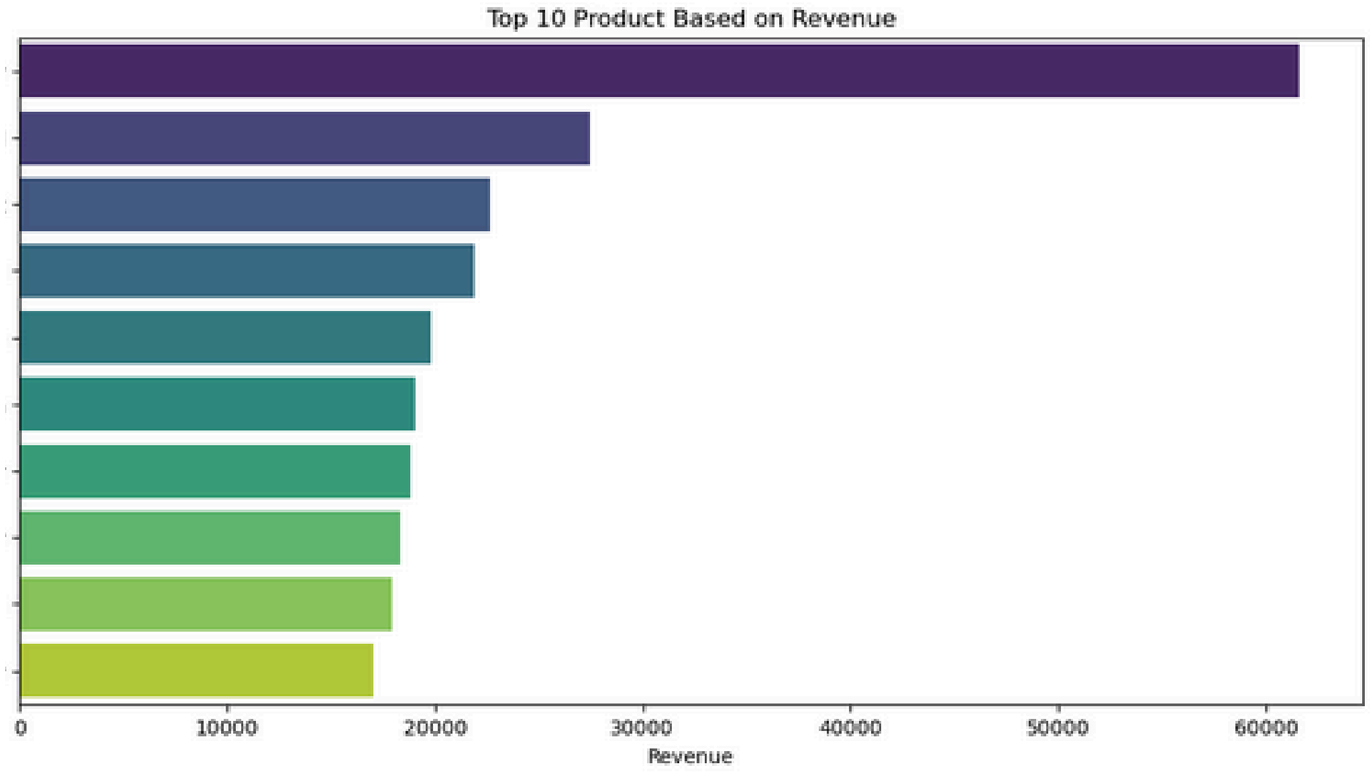


	Order month	Sales
0	1	94924.8356
1	2	59751.2514
2	3	205005.4888
3	4	137762.1286
4	5	155028.8117
5	6	152718.6793
6	7	147238.0970
7	8	159044.0630
8	9	307649.9457
9	10	200322.9847

- Lowest Sales: February (~59,751) → Possible post-holiday effect.
- Highest Sales: November (~307,649) → May be affected by major shopping events (Year End Sales).
- Significant spike in March, then stable trend until August.
- Increase starts in September, peaks in November, then slightly drops in December.
- The company needs a more aggressive promotion strategy at the beginning of the year and stock optimization in peak sales months.

Top 10 Products Based on Revenue

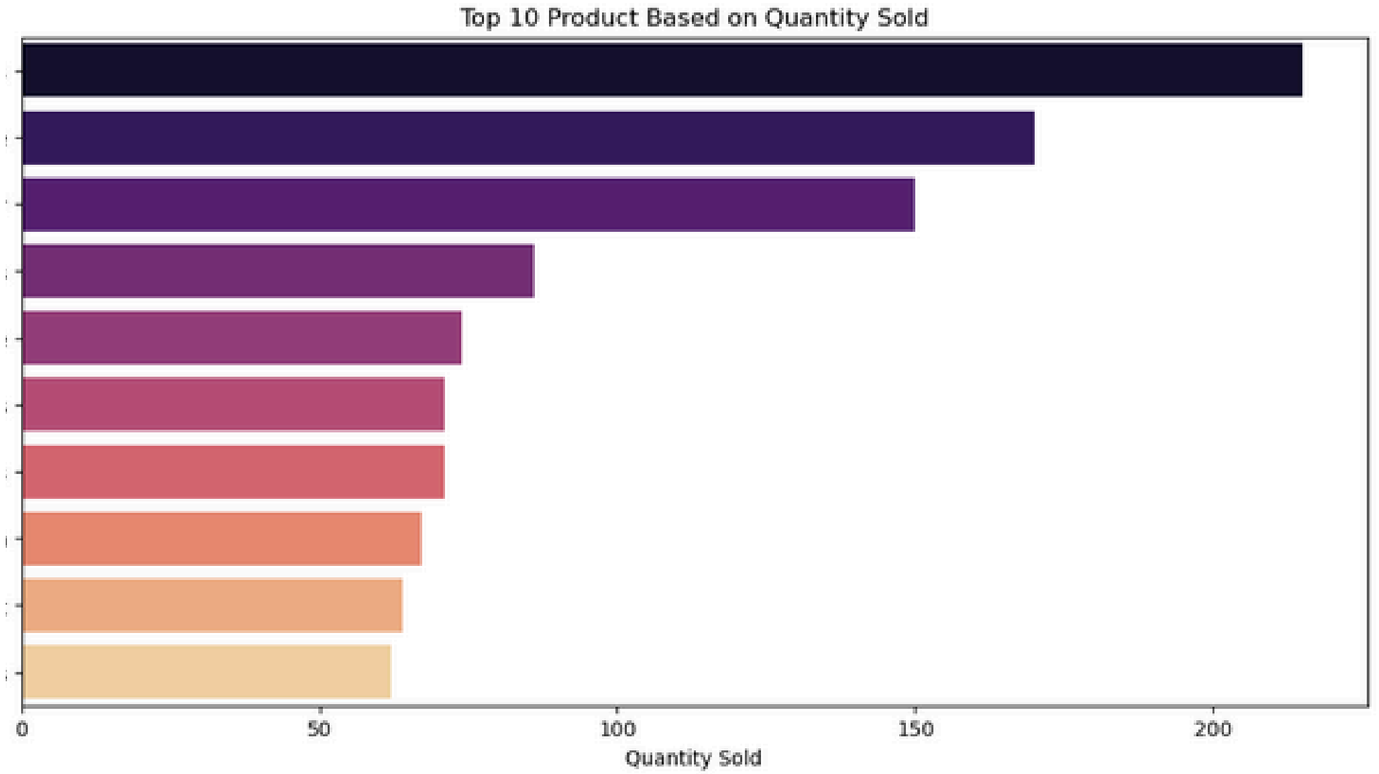
	Product Name	Sales
404	Canon imageCLASS 2200 Advanced Copier	61599.824
650	Fellowes PB500 Electric Punch Plastic Comb Bin...	27453.384
444	Cisco TelePresence System EX90 Videoconferenci...	22638.480
786	HON 5400 Series Task Chairs for Big and Tall	21870.576
686	GBC DocuBind TL300 Electric Binding System	19823.479
688	GBC Ibimaster 500 Manual ProClick Binding System	19024.500
805	Hewlett Packard LaserJet 3310 Copier	18839.686
787	HP Designjet T520 Inkjet Large Format Printer ...	18374.895
683	GBC DocuBind P400 Electric Binding System	17965.068
813	High Speed Automatic Electric Letter Opener	17030.312



- Highest Sales: Canon imageCLASS 2200 Advanced Copier (~61,599)
- Lowest Sales: High Speed Automatic Electric Letter Opener (~17,030).
- Copiers and binding systems dominated the sales which is a high demand for office equipment.

Top 10 Products Based on Quantity Sold

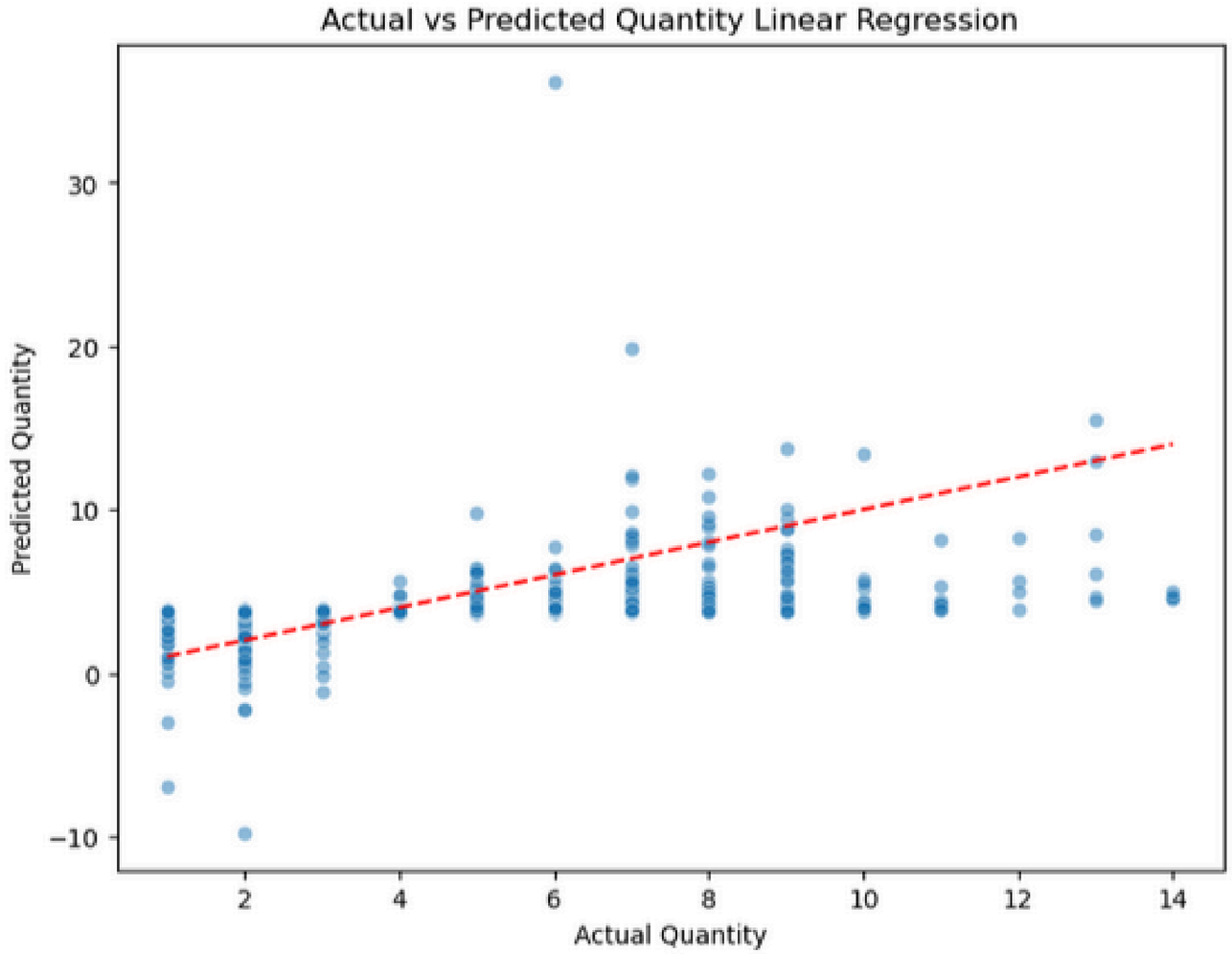
	Product Name	Quantity
1499	Staples	215
1493	Staple envelope	170
537	Easy-staple paper	150
1500	Staples in misc. colors	86
942	KI Adjustable-Height Table	74
259	Avery Non-Stick Binders	71
1511	Storex Dura Pro Binders	71
699	GBC Premium Transparent Covers with Diagonal L...	67
1450	Situations Contoured Folding Chairs, 4/Set	64
1497	Staple-based wall hangings	62



- Highest Quantity Sold: Staples (215 units)
- Lowest Quantity Sold: Staple-based wall hangings (62 units)
- Products with high quantity sold are generally basic office supplies (staples, paper, envelopes).

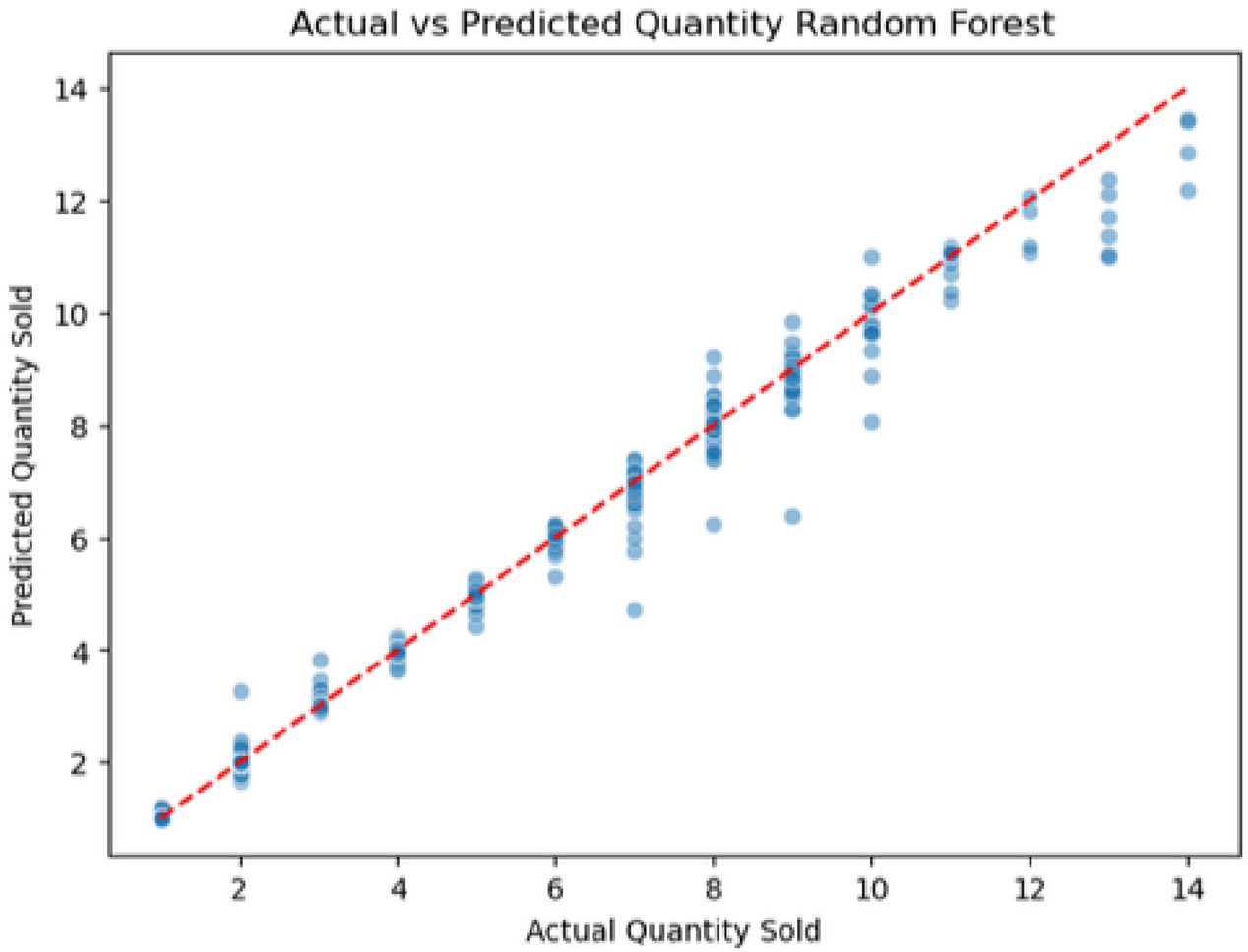
Feature vs Quantity Sold

Feature	Coefficient	Interpretation
Discount	0.060222	Discounts have the most positive influence on sales. The bigger the discount, the more items are sold.
Order Month	0.012535	The month of ordering also matters, meaning there is a seasonal trend in sales.
Order Year	0.009257	Ordering years have a slight effect, there could be growth or changes in trends from year to year.
Sales	0.003850	Total sales slightly affects the number of items sold, but not significantly.
Category	-0.003340	Product categories have a negative influence, meaning that certain categories can reduce the number of items sold.
Price	-0.014489	Price has the largest negative effect, meaning that the more expensive the item, the less quantity is sold (logical due to the demand vs price effect).



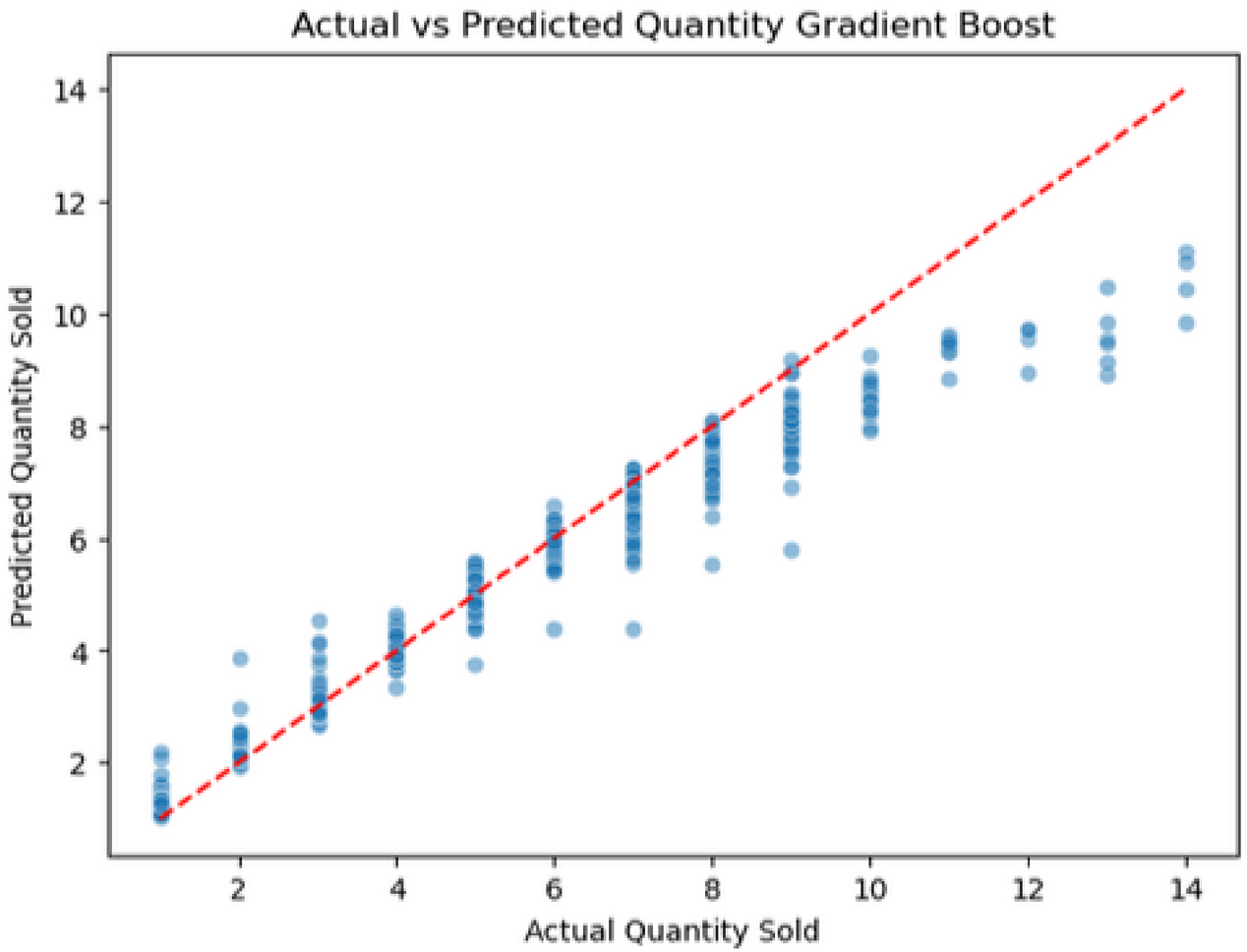
Quantity Sold Prediction with Linear Regression

- The scatter of points far enough away from the regression line indicates considerable error.
- The model tends to be less accurate for predicting quantities at higher values.



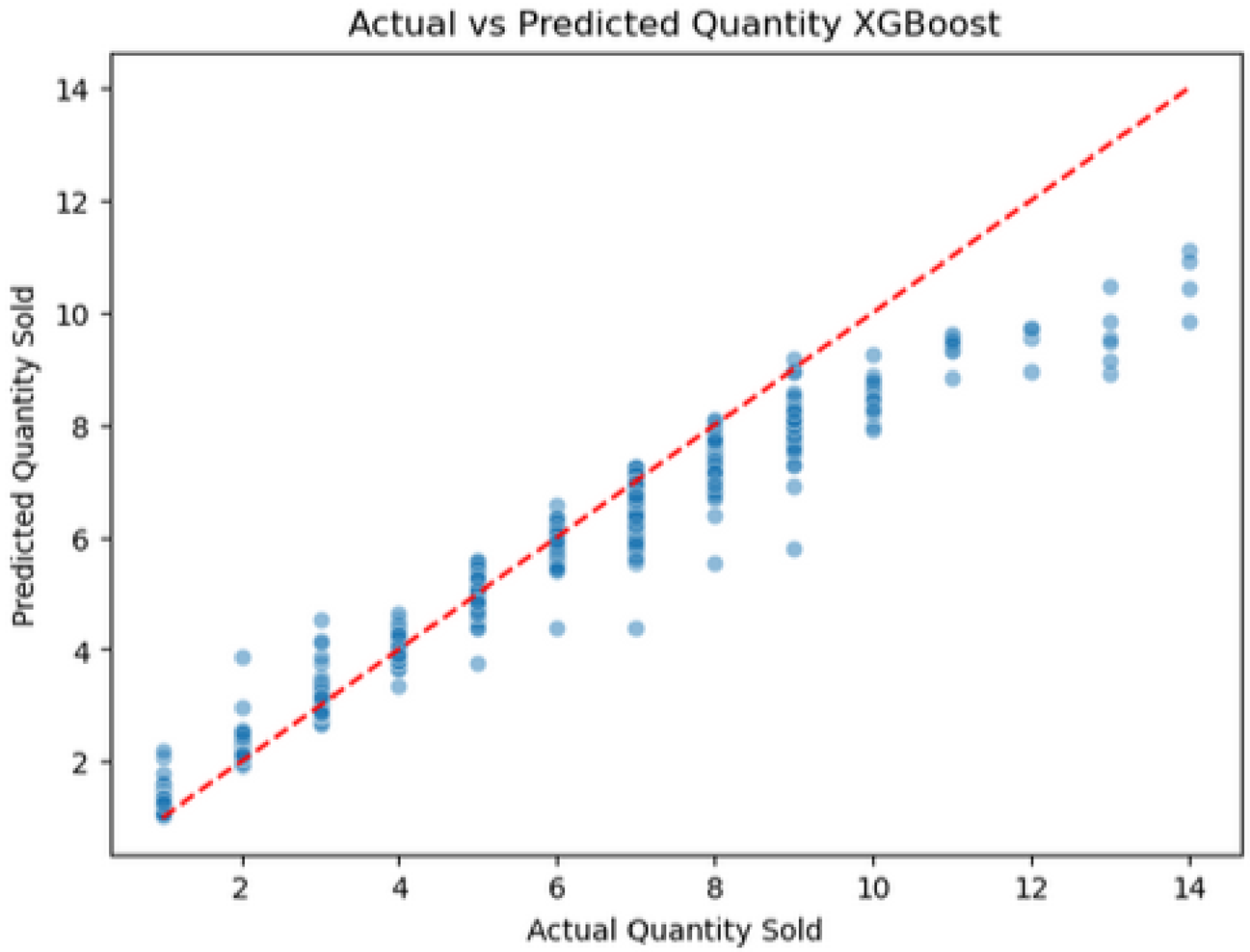
Quantity Sold Prediction with Random Forest

- The distribution of points is tighter around the line than the Linear Regression model, indicating a smaller error.
- This model is better at capturing patterns than the previous model, making it more suitable for sales quantity prediction.



Quantity Sold Prediction with Gradient Boost

- The Gradient Boost model shows a fairly good relationship between the actual and predicted quantities.
- Compared to Random Forest, this model tends to be more spread out, indicating that there is slightly more error.



Quantity Sold Prediction with XGBoost

- The XGBoost model shows an excellent prediction pattern with the distribution of points following the ideal line (red line).
- The prediction results are more accurate than Linear Regression and Gradient Boost, with less deviation.
- Overall, XGBoost is one of the best models for sales quantity prediction in this dataset.

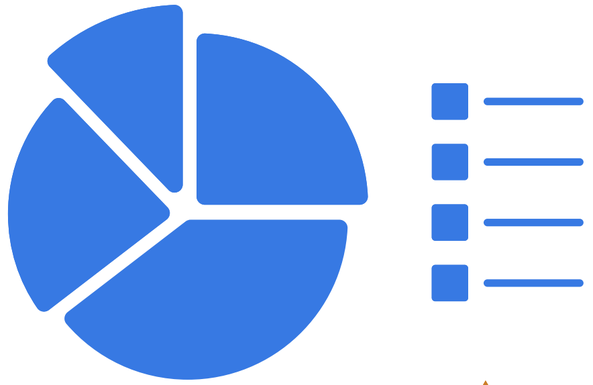
Model Evaluation Result

Model	MAE	RMSE	R^2 Score
Linear Regression	1.50	2.10	0.13
Random Forest	0.05	0.18	0.99
Gradient Boost	0.27	0.47	0.96
XGBoost	0.10	0.21	0.99

- Random Forest and XGBoost have the best performance, with low MAE & RMSE and R^2 close to 1.
- Linear Regression is less suitable, as the R^2 is only 0.13, indicating this model cannot capture data patterns well.
- Gradient Boosting is quite good, but still inferior to Random Forest & XGBoost in terms of error and accuracy.

Conclusion

- The highest sales occurred in November which was likely influenced by major events such as the Year End Sale, while the lowest sales were recorded in February, likely due to the post-holiday effect. The product with the highest sales value was the Canon imageCLASS 2200 Advanced Copier, indicating the high demand for office equipment such as copiers and binding systems. Meanwhile, the product with the highest quantity was staples, reflecting the regular need for basic office supplies.
- In terms of prediction modeling, Random Forest and XGBoost provide the best performance with low MAE and RMSE values and R^2 close to 1. In contrast, Linear Regression shows less optimal results because it is unable to capture data patterns well.



Data Analyst Project
2025

Thank You!

Muhammad Fakhri Azhar