Supermarket Sales

Data Science Project

by

Walaa Nasr Elghitany

30-4-2024

Project Overview

Objective: A supermarket chain aims to optimize its sales and inventory management by accurately predicting future sales for each product across its various stores. we want to develop a machine learning model to predict supermarket sales based on historical sales data and other relevant factors.

Data Understanding

Data describing:

1000 entries with total 17 columns:

Invoice ID, Branch, City, Customer type, Gender, Product line, Unit price, Quantity, Tax 5%, Total, Date ,Time, Payment, cogs, gross margin percentage, gross income, & Rating

this dataset collected in 2019 during 1st 3 months of the year, january, february and march

collected from only 3 branches in 3 cities (Yangon, Mandalay, Naypyidaw) with 6 categories of product lines (Fashion accessories, Food and beverages, Electronic accessories, Sports and travel, Home and lifestyle, Health and beauty)

only 2 customer types (member and normal) with 3 types of payment (Ewallet, Cash, Credit card).

Strengths and Limitations

This dataset has a couple of strengths and weaknesses that will be highlighted. First of all, one **strength** of this dataset is its size: it contains approximately 1000 rows and can therefore be seen as a relatively big dataset. Another strength is that the sales data is structured, and therefore it is easily understandable.

As for **limitations**, it depends on the exact predictions that will be made based on the data. some irrelevant columns can be challenging to detect and filter out. all dataset is equally divided between all categories with vey slight differences as if it is generated data and not real one. there are some columns that totally depends on each others like: (Unit price *Quantity) + Tax 5% = Total

Data Collection and Preprocessing

Data Sources: supermarket sales, send by skill genie.

Data Preparation: Before we can start creating our data models, it is necessary to prepare the data in the right way. The models require the data to be in a form differently from how it is provided in its natural, raw way. Therefore, some conversion of our dataset is necessary.

Data Cleaning:

removing duplicates: no duplicates were found
handling missing values: no missing data
addressing outliers: we only take a look at them

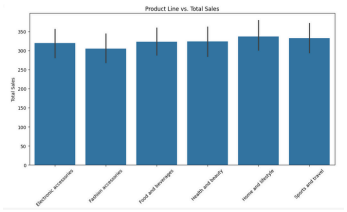
Data Analysis & visualization:

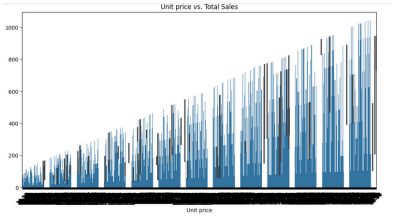
- **Descriptive Statistics: Provide basic statistics like mean, median, mode, range, and standard deviation for relevant columns.**

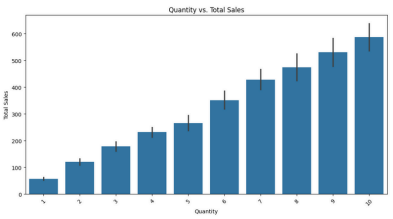
	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	0.000000	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905	49.650000	10.00000

Data Analysis & Visualization:

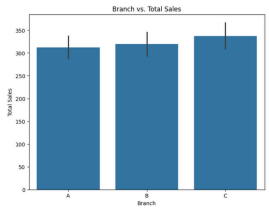
- **Product Analysis: sales performance by Product Attributes:**
  - Product line
  - Unit price
  - Quantity

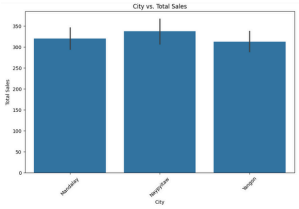




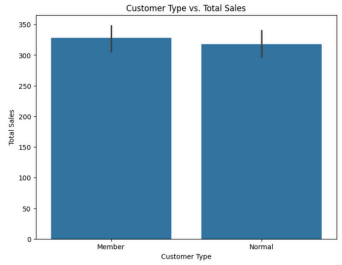


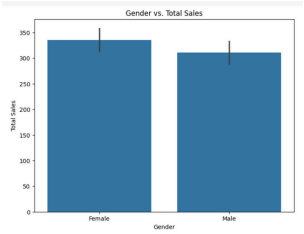
- **Store Location: Sales by product category, and sales performance by**
  - Branch
  - City

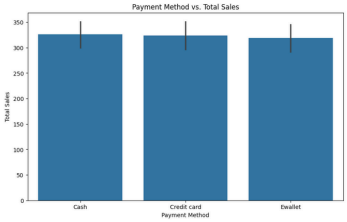




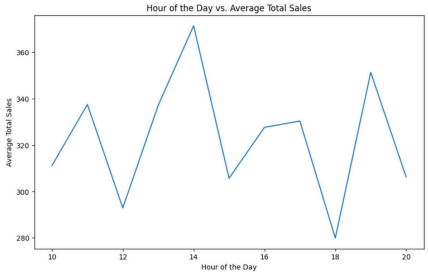
- **Customer Analysis:** sales performance by customer demographics, purchasing patterns, loyalty program effectiveness.
  - type
  - gender
  - payment

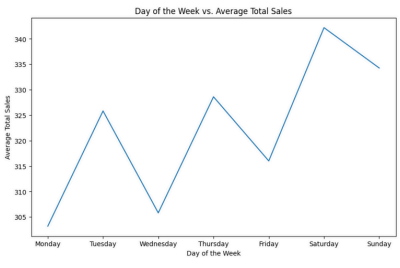


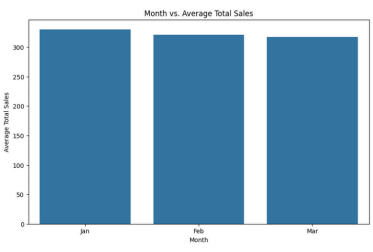




- **Sales Analysis:** Analyze total sales, sales trends over time







Feature Engineering

Feature Creation:

- we extracted these columns from date and time columns : DayOfWeek','Month','Hour'
- 'Product line' and sum up the 'Total' column

	Product line	TotalRevenue
0	Electronic accessories	54337.5315
1	Fashion accessories	54305.8950
2	Food and beverages	56144.8440
3	Health and beauty	49193.7390
4	Home and lifestyle	53861.9130
5	Sports and travel	55122.8265

- Average Transaction Amount per Hour

	Hour	AvgTransactionAmount
0	10	311.103772
1	11	337.525883
2	12	292.875084
3	13	337.118709
4	14	371.426494
5	15	305.681456
6	16	327.614591
7	17	330.340784
8	18	279.896129
9	19	351.323124
10	20	306.260360

- Total Sales per Product and Gender

	Product line	Gender	TotalSales
0	Electronic accessories	Female	27102.0225
1	Electronic accessories	Male	27235.5090
2	Fashion accessories	Female	30437.4000
3	Fashion accessories	Male	23868.4950
4	Food and beverages	Female	33170.9175
5	Food and beverages	Male	22973.9265
6	Health and beauty	Female	18560.9865
7	Health and beauty	Male	30632.7525
8	Home and lifestyle	Female	30036.8775
9	Home and lifestyle	Male	23825.0355
10	Sports and travel	Female	28574.7210
11	Sports and travel	Male	26548.1055

- Total Sales per Customer and Payment Method:

	Customer type	Payment	TotalSales
0	Member	Cash	54661.0155
1	Member	Credit card	57771.4725
2	Member	Ewallet	51790.9560
3	Normal	Cash	57545.5545
4	Normal	Credit card	42995.5995
5	Normal	Ewallet	58202.1510

- Average Rating per Product and Branch

	Product line	Branch	AvgRating
0	Electronic accessories	A	6.911667
1	Electronic accessories	B	7.116364
2	Electronic accessories	C	6.747273
3	Fashion accessories	A	6.878431
4	Fashion accessories	B	6.722581
5	Fashion accessories	C	7.440000
6	Food and beverages	A	7.253448
7	Food and beverages	B	6.994000
8	Food and beverages	C	7.080303
9	Health and beauty	A	6.900000
10	Health and beauty	B	7.100000
11	Health and beauty	C	6.998077
12	Home and lifestyle	A	6.930769
13	Home and lifestyle	B	6.516000
14	Home and lifestyle	C	7.060000
15	Sports and travel	A	7.257627
16	Sports and travel	B	6.509677
17	Sports and travel	C	7.028889

Data Transformation & Encoding

- Data was subdivided into numerical data and we changed data and time columns into datatime type
- Data encoding of categorical columns using one hot encoder
- we deleted redundant data like, Invoice ID','Time','Year

• Data preprocessing

data splitting: we split the data using train-test split
data scaling: we scaled data using StandardScaler
our Data is **balanced**, so we don't need to do any further process.

Model Selection and Development

Model Candidates:
Initial Set of Models Considered:
- **Logistic Regression:** Chosen for its efficiency and effectiveness in binary classification tasks, especially useful if the goal is to predict binary outcomes (e.g., predicting whether a customer will buy a specific product or not).
-**K-Nearest Neighbors (KNN):** Selected due to its simplicity and effectiveness in capturing patterns in data without making assumptions about the underlying data distribution, useful in scenarios where relationships might be non-linear.
- **Random Forest:** Known for its high accuracy and robustness, capable of handling large data sets with higher dimensionality without overfitting, making it suitable for a dataset with diverse features and complex relationships.

2. Model Comparison
Evaluation Metrics:

- Models were primarily evaluated using Mean Squared Error (MSE). These metrics help in understanding the average error magnitude and the consistency of error terms across data predictions.
- R^2 Score, or the coefficient of determination, was also considered to gauge how well future samples are likely to be predicted by the model.

•

Cross-Validation:
Employed k-fold cross-validation (k=5) to validate the performance of each model across different subsets of the data, ensuring the model's ability to perform consistently across various unseen datasets.

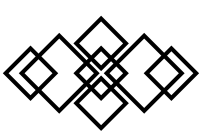
- **Performance Analysis:**
  - **Linear Regression:**
    - Showed the lowest MSE and highest R^2 values indicating strong predictive performance with minimal error variance.
    - Its simplicity and fewer assumptions about data distribution may have contributed to a robust performance across different data splits.
  - **KNN:**
    - Did not perform as well, potentially due to improper scaling of features or suboptimal choice of 'k'. KNN's performance can deteriorate if the distance metric is skewed by features varying in scale.
  - **Random Forest Regression:**
    - While typically strong, it might not have reached its potential due to overfitting, especially if hyperparameters (like tree depth and the number of trees) were not optimally tuned.
- **Final Model Choice**

Rationale for Selecting Linear Regression:

- Efficiency and Effectiveness: Despite its simplicity, Linear Regression provided the most accurate predictions for total sales, suggesting that the relationship between the variables and the sales was predominantly linear.
- Interpretability: Offers clear interpretation of model outputs, which is highly valued in business settings for making informed decisions.
- Computational Simplicity: Requires less computational resources compared to KNN or Random Forest, making it scalable and easier to deploy.

Conclusion

- The choice of Linear Regression as the final model was driven by its superior performance on key metrics, indicating effective handling of the linear patterns in the dataset. Its simplicity and interpretability make it an excellent tool for stakeholders who need to understand and act on the predictions.



Key Insights from Model Results

1. **Uniform Distribution Across Categories**
   - **Balanced Data:** The dataset exhibits an equal distribution of sales across all branches, product lines, and customer demographics. This balance provides a unique opportunity to analyze supermarket operations without the confounding effects of unequal sample sizes.
   - **Implications for Analysis:** The uniformity in the dataset ensures that insights derived are not biased toward any particular group or category, enhancing the reliability of the findings.
2. **Consistent Performance Across Payment Methods**
   - **Equal Use of Payment Options:** There is an even split in payment methods (Ewallet, Cash, Credit Card), suggesting flexibility and customer adaptability in payment preferences.
   - **Effect on Sales Strategy:** The absence of a dominant payment method may indicate the need for maintaining diverse payment options to cater to all customer preferences.
3. **General Customer Satisfaction**
   - **Stable Rating Distribution:** Customer ratings are well-distributed, indicating general satisfaction across the product lines and store branches. This suggests consistent service quality and customer experience across the board.
   - **Business Stability:** Uniform customer ratings across different times and dates reinforce the notion of operational consistency and reliability in customer service.

Data Insights

1. **No Significant Seasonal or Temporal Trends**
   - **Constant Sales Volume:** Analysis reveals no significant fluctuations in sales volume across different times of the year or specific dates, which is unusual for retail but indicates a stable demand throughout.
   - **Strategic Planning:** The lack of seasonal trends simplifies inventory and staffing management but also suggests the potential for strategic promotions to boost sales during typically slower periods.
2. **Equal Representation Across Customer Types and Genders**
   - **Demographic Distribution:** Both genders and customer types (Member, Normal) are equally represented, providing a comprehensive view of consumer behavior across different demographic segments.

- **Marketing Strategies:** Marketing and sales strategies can be universally applied without the need for customization based on gender or membership status, simplifying marketing efforts.

3. **Product Line Performance**

- E**ven Product Sales:** Each product line (e.g., Health and Beauty, Sports and Travel) contributes equally to the total sales, which is rare and indicates a well-rounded product offering.
- **Inventory Management:** Equally strong performance across product lines allows for predictable inventory management and reduces the risk associated with overstocking specific items.

Conclusion

The balanced nature of the dataset across various facets provides a robust platform for analyzing business operations without the need to account for typical biases seen in retail data. This allows for straightforward application of findings to improve business practices and customer satisfaction. Future analyses might focus on experimenting with targeted promotions or varied marketing strategies to test for potential improvements in sales dynamics, given the stable baseline established by the current data characteristics.

Recommendations for future improvements:

Feature Suggestions

1. **Demographic Data Expansion:**

- Age and Income Levels: Include customer demographic details such as age and income, which could help in better understanding purchasing patterns and tailoring marketing strategies.
- Loyalty Programs:** Track participation in loyalty programs more explicitly to analyze its impact on sales and customer retention.

2. **External Data Integration:**

- **Weather Data:** Integrate local weather data to examine its impact on sales of specific product lines like seasonal goods.
- **Economic Indicators:** Include local economic indicators such as employment rates or consumer confidence indices that could affect consumer spending behaviors.

3.

3**- Operational Metrics:**
- **Stock Levels:** Incorporate inventory levels to help predict stock outs and optimize stock replenishment cycles.
- **Store Traffic:** Monitor foot traffic data within stores to better align staffing needs and promotional activities.

Operational Integration

1. **Real-Time Analytics:**
- Dashboard Implementation: Develop interactive dashboards to provide real-time insights into sales trends, inventory levels, and customer feedback, allowing for quicker managerial decisions.
- **Alert Systems:** Set up automated alerts for unusual sales activity or inventory levels to immediately address potential issues.

2. **Continuous Learning and Adaptation:**
- **Feedback Loop:** Implement a feedback system where the model's predictions and their accuracy are regularly reviewed to refine the algorithms based on new data and changing market conditions.
- **A/B Testing:** Regularly conduct A/B testing for different sales strategies based on model recommendations to continuously improve and adapt business strategies.

By addressing these areas, the supermarket can not only improve the accuracy and relevance of its predictive models but also enhance operational efficiency and responsiveness to changing market dynamics.

This approach ensures that the supermarket remains competitive and can adapt quickly to new opportunities and challenges.

Summary

This project set out to analyze supermarket sales data using various predictive models with the primary objective of understanding and forecasting sales dynamics. Through careful data preprocessing, model selection, and analysis, we discovered that the dataset was remarkably balanced across various categories and subsets. Linear Regression emerged as the most effective model, providing valuable insights into the factors driving sales, despite the simplicity of the model and the complexity of the dataset.

The analysis revealed key insights such as the uniform distribution of sales across different branches and customer demographics, consistent performance across payment methods, and general customer satisfaction. The stability and lack of significant trends in the data suggested a steady market environment, providing a reliable foundation for operational and strategic decision-making.