Supermarket Sales Patterns and Predictions

Akash Kumar Kondaparthi
Dept. of Electrical and Computer Engineering
University of Florida
Gainesville, FL, USA
akash.kondaparth@ufl.edu

***Abstract*—In the contemporary retail environment, data-driven insights are important to drive business strategies and make informed decisions. This project delves into the sales data of a supermarket chain to uncover insights into understanding consumer behavior, sales patterns, and influential factors by utilizing simple machine-learning techniques to address pivotal questions related to sales, customer preferences, and potential revenue-maximization strategies. Further, intricate relationships between variables such as product lines, payment methods, time slots, and gender were elucidated using interaction terms and logistic regression. The findings not only shed light on consumer behavior in supermarket contexts but also provide a blueprint for retailers to optimize their marketing, stocking, and operational strategies for maximal profitability based on factors affecting the respective sales.**

I. Introduction

Businesses need to harness the data to gain insights into consumer behavior, optimize sales strategies, and improve operational efficiencies. This project aimed to provide insights into the factors affecting sales, the relationship between various product attributes, and the prediction of customer behavior for a supermarket. This project is centered around the dataset involving sales transactions, where the aim is multifaceted: from predicting gross income to decoding factors influencing unit prices, gender preferences, and customer types.

While this project is prediction-oriented, it also aids in interpreting the models to deduce the relationship and significance of various features in prediction. Using methods like coefficient plots and stem plots, the magnitude and direction of relationships between predictors and outcomes were elucidated. Moreover, metrics such as R-squared, MAE, and MSE were used to paint a clear picture of the model performance, along with confidence intervals.

II. Data Preprocessing

The dataset sourced for this analysis encompasses sales transactions of a supermarket. Critical attributes include 'Date', 'Time', 'Gender', 'Customer Type', 'Product Line', 'Unit Price', 'Quantity', 'Payment', and 'gross income'.

The complete list of data preprocessing steps undertaken includes:

1. Feature Exclusion: 'Invoice ID' was identified as a unique identifier. So, it was removed from the dataset.
2. Handling Redundancy: There was an overlying redundancy between 'Branch' and 'City'. A clear one-to-one mapping was evident, with each branch corresponding uniquely to a city. That is, Branch A = Yangon, Branch B = Mandalay, and Branch C = Naypyitaw. The project chose to retain 'Branch' and drop 'City'.
3. Encoding Features: Certain features possessed a natural order, such as 'Date' and 'Time'. Here, label encoding was adopted, converting these features into a series of integers. For attributes with binary values, like 'Gender' and 'Customer type', label encoding was again favored over One-Hot Encoding, ensuring that the feature space remained efficient without inflating the number of columns.
4. Feature Scaling: Feature scaling was needed since there are scale-sensitive models like logistic regression being used. Standardization ensured that no particular feature dominated the models due to its scale, leading to a more balanced and fair prediction mechanism.
5. Data Splitting: To ensure robust evaluation, the data was split into training and testing sets before the beginning of data preprocessing. This allowed models to be trained on one subset of the data and then validated on an unseen set, and most importantly, avoided data leakage.

III. Training the Models

3.1. Predicting Gross Income:

A. Linear Regression: Linear regression was used to predict the 'gross income' from the training data. The entire process is encapsulated in a pipeline. Upon training the model with the training data, its performance has been evaluated using various regression metrics:

- R-squared (R2) Score: This metric provides the proportion of the variance in the dependent variable that is predictable from the independent variables. For the model, R2 was found to be 1.0.

- Mean Absolute Error (MAE) and Mean Squared Error (MSE): These metrics provide a measure of the prediction error. For the trained model, the errors were negligible. Additionally, a 95% Confidence Interval for the R2 score was estimated to be [1, 1].

B. Lasso Regression: Lasso Regression is a type of linear regression that includes a regularization term. The regularization term discourages overly complex models which can overfit the training data. Lasso regression with 5-fold cross-validation was implemented to optimize the alpha. The R2 score for Lasso Regression was found to be 0.99, which is slightly less than the Linear Regression. The 95% Confidence Interval for the R2 score was [0.98994155, 0.98999999].

3.2. Predicting Unit price: A similar methodology has been followed for predicting Unit Price.

A. Linear Regression: For predicting 'Unit price', a linear regression model was trained. The R2 score was 0.7784, MAE was 0.3414, and MSE was 0.2215. The 95% Confidence Interval for the R2 score was [0.75521218, 0.80064739] obtained on the training data.

B. Lasso Regression: The Lasso Regression model for 'Unit price' prediction had an R2 score of 0.5049, with MAE and MSE values being 0.5739 and 0.4951 respectively. The

95% Confidence Interval for the R2 score was [0.47077485, 0.53700979].

3.3. Classifying Gender using Logistic Regression:

Logistic regression with polynomial interaction terms (degree 2) was employed to classify gender. After training, the model achieved an accuracy of 0.75 on the training data, and the R2 score was 0.0218 with a 95% confidence interval between -0.0054 and 0.0489.

3.4. Classifying Customer type using Logistic Regression:

A similar logistic regression model with polynomial interaction terms was used for this task. The best hyperparameters found for the model were a regularization strength (C) of 0.1, L2 penalty, and 'liblinear' solver. The model's accuracy on the training data was 0.69, and the R2 score was 0.0200, with a 95% confidence interval ranging from -0.01399 to 0.05402.

3.5. Predicting the day of purchase:

A. Logistic Regression: The logistic regression model was used to predict the 'day of purchase' with a similar methodology followed to classify Gender and Customer type. The model was trained with various hyperparameters for regularization and solvers using Grid Search. The obtained R^2 score was 0.2284.

B. Random Forest Classifier

The Random Forest Classifier model was also implemented with the same methodology using Grid Search to fine-tune the hyperparameters. The obtained R^2 score was 0.7869. All these models were trained into a pickle (.pkl) file for later use in testing on the test set.

IV. Testing and Results

The testing has been done on a separate test set that has been separated from the whole dataset before data preprocessing to avoid data leakage during standardization. During testing, the trained models have been retrieved from the pickle files to test them on the test data. The following are the obtained results in the training phase:

4.1. Predicting 'Gross Income' using Regression Models

A. Linear Regression:

- R-squared: 1.0, MAE: 8.765e-16, and MSE: 1.165e-30

- 95% Confidence Interval for R^2 Score: [1.0, 1.0]

B. Lasso Regression:

- R2 Score: 0.989, MAE: 0.086 and MSE: 0.0109

- 95% Confidence Interval for R^2 Score: [0.9894, 0.9899]

4.2. Predicting 'Unit Price' using Regression Models

A. Linear Regression:

- R-squared: 0.792, MAE: 0.354 and MSE: 0.2154

- 95% Confidence Interval for R^2 Score: [0.7476, 0.8308]

B. Lasso Regression:

- R2 Score: 0.491, MAE: 0.615 and MSE: 0.5288

- 95% Confidence Interval for R^2 Score: [0.4101, 0.5552]

4.3. Classification for 'Gender' using Logistic Regression:

- R^2 (coefficient of determination): 0.0850

- 95% Confidence Interval for R^2: (-0.0396, 0.2096)

	precision	recall	f1-score	support
0.0	0.62	0.67	0.64	42
1.0	0.50	0.45	0.47	31
accuracy			0.58	73
macro avg	0.56	0.56	0.56	73
weighted avg	0.57	0.58	0.57	73

R^2 (coefficient of determination): 0.0850
95% confidence interval for R^2: (-0.0396, 0.2096)

Figure 1. Classification report of Logistic Regression predicting Gender.

4.4. Classification for 'Customer Type' using Logistic Regression:

- R^2: 0.0639 and 95% CI for R^2: (0.00313, 0.124698)

	precision	recall	f1-score	support
0.0	0.70	0.59	0.64	44
1.0	0.50	0.62	0.55	29
accuracy			0.60	73
macro avg	0.60	0.61	0.60	73
weighted avg	0.62	0.60	0.61	73

Figure 2. Classification Report for Logistic Regression predicting Customer type.

4.5. Classification for 'Day of Purchase' using Logistic Regression

- R^2: -0.2284, and 95% CI for R^2: [-0.38357, -0.11651]

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	26
1.0	0.00	0.00	0.00	36
2.0	0.14	1.00	0.25	28
3.0	0.00	0.00	0.00	25
4.0	0.00	0.00	0.00	29
5.0	0.00	0.00	0.00	28
6.0	0.00	0.00	0.00	28
accuracy			0.14	200
macro avg	0.02	0.14	0.04	200
weighted avg	0.02	0.14	0.03	200

Figure 3. Classification Report for Logistic Regression predicting Date

4.6. Classification for 'Day of Purchase' using Random Forest Classifier:

- R^2: -0.651, and 95% CI for R^2: [-0.9669, -0.3925]

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	26
1.0	0.12	0.03	0.05	36
2.0	0.11	0.18	0.14	28
3.0	0.12	0.12	0.12	25
4.0	0.15	0.17	0.16	29
5.0	0.20	0.43	0.27	28
6.0	0.30	0.21	0.25	28
accuracy			0.16	200
macro avg	0.14	0.16	0.14	200
weighted avg	0.15	0.16	0.14	200

Figure 4. Classification report for Random Forest Classifier predicting Date

V. Observations

From the training and testing phases with the data, the following observations have been noted:

1. When predicting Gross income, it is noted that the following features were excluded by Lasso, which means that these features did not contribute to predicting gross income: 'Date', 'Time', 'Gender', 'Customer type', 'Unit price', 'Quantity', 'gross margin percentage', 'Rating', 'Branch', 'Product line', 'Payment'. And, the following features have been included by Lasso: 'Total', and 'cogs', meaning these features contributed to predicting gross income.

   While training a Lasso regressor and a Logistic regressor the following coefficients were obtained:

	Feature	Linear Regression Coeff	Lasso Regression Coeff
0	Date	2.428510e-17	0.000000e+00
1	Time	3.442127e-17	-0.000000e+00
2	Gender	2.414470e-16	-0.000000e+00
3	Customer type	2.785810e-16	0.000000e+00
4	Unit price	-3.602013e-16	0.000000e+00
5	Quantity	7.894451e-18	0.000000e+00
6	Total	5.000000e-01	9.000000e-01
7	cogs	5.000000e-01	7.105427e-17
8	gross margin percentage	-5.551115e-17	0.000000e+00
9	Rating	-3.652268e-16	-0.000000e+00
10	Branch_A	-2.944968e-17	-0.000000e+00
11	Branch_B	-1.537721e-16	-0.000000e+00
12	Branch_C	1.187090e-16	0.000000e+00
13	Product line_Electronic accessories	2.357093e-17	0.000000e+00
14	Product line_Fashion accessories	2.334060e-17	-0.000000e+00
15	Product line_Food and beverages	-9.181281e-17	-0.000000e+00
16	Product line_Health and beauty	-3.116601e-17	0.000000e+00
17	Product line_Home and lifestyle	4.744297e-17	0.000000e+00
18	Product line_Sports and travel	7.056775e-17	-0.000000e+00
19	Payment_Cash	-1.173943e-16	-0.000000e+00
20	Payment_Credit card	-4.227018e-17	0.000000e+00
21	Payment_Ewallet	1.617653e-16	-0.000000e+00

Figure 5. Coefficients Table for predicting gross income

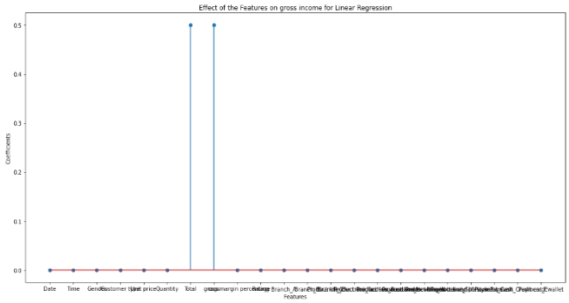

Figure 6. Stem plot of the feature coefficients of Linear Regression

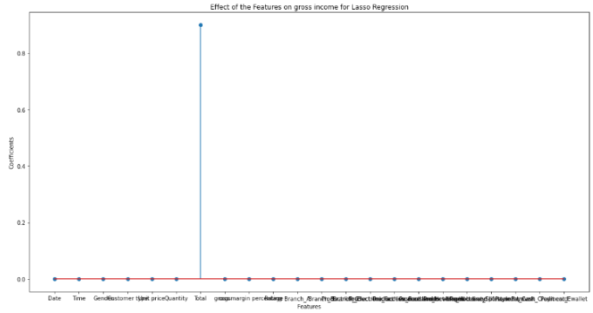

Figure 7. Stem plot of the feature coefficients of Lasso Regression

Hence, from the plots, features: 'Date', 'Time', 'Gender', 'Customer type', 'Unit price', 'Quantity', 'gross margin percentage', 'Rating', 'Branch', 'Product line', and 'Payment' is excluded by the Lasso regressor. This happened because Lasso promotes sparsity and possibly these features provided less information in predicting gross income. Only the features 'Total' and 'cogs' contributed to predicting 'gross income'. The stem plot shows the importance of these features in predicting gross income. The effect of variables like unit price, and quantity on the gross income can be interpreted from the coefficients table. Features with higher absolute values have a greater impact on the gross income. A positive coefficient indicates a direct relationship while a negative coefficient indicates an inverse relationship. So from the table, for example, according to the Lasso Regressor, the feature Total has a higher influence on gross income than cogs.

2. When predicting unit price, again, the effect of variables on the unit price can be interpreted from the coefficients table or the stem plot shown. Features with higher absolute values have a greater impact on the gross income. A positive coefficient indicates a direct relationship while a negative coefficient indicates an inverse relationship.

	Feature	Linear Regression Coeff	Lasso Regression Coeff
0	Date	-0.004792	-0.000000e+00
1	Time	0.021095	0.000000e+00
2	Gender	0.019647	0.000000e+00
3	Customer type	-0.067488	-0.000000e+00
4	Quantity	-0.857272	-1.840399e-01
5	Total	0.414049	5.638319e-01
6	cogs	0.414049	3.197442e-16
7	gross margin percentage	0.000000	0.000000e+00
8	gross income	0.414049	0.000000e+00
9	Rating	0.015859	-0.000000e+00
10	Branch_A	0.011617	-0.000000e+00
11	Branch_B	0.002248	0.000000e+00
12	Branch_C	-0.013865	0.000000e+00
13	Product line_Electronic accessories	-0.000937	-0.000000e+00
14	Product line_Fashion accessories	0.025453	0.000000e+00
15	Product line_Food and beverages	0.021996	0.000000e+00
16	Product line_Health and beauty	0.000198	-0.000000e+00
17	Product line_Home and lifestyle	-0.037784	-0.000000e+00
18	Product line_Sports and travel	-0.008927	-0.000000e+00
19	Payment_Cash	0.034596	0.000000e+00
20	Payment_Credit card	-0.049414	-0.000000e+00
21	Payment_Ewallet	0.014818	-0.000000e+00

Figure 8. Coefficients Table for predicting Unit price

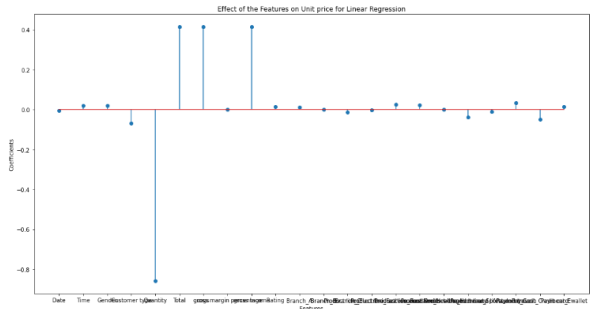

Figure 9. Stem plot of the feature coefficients of Linear Regression for Unit Price

So from the table (or the stem plot), for example, according to the Lasso Regressor, the feature Total has a higher positive influence on unit price and Quantity has a higher negative effect on unit price.

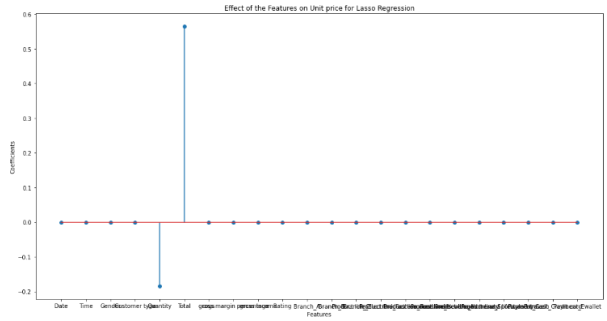

Figure 10. Stem plot of the feature coefficients of Lasso Regression for Unit Price

3. When classifying Gender, the following observations have been made using degree two interactions among features:

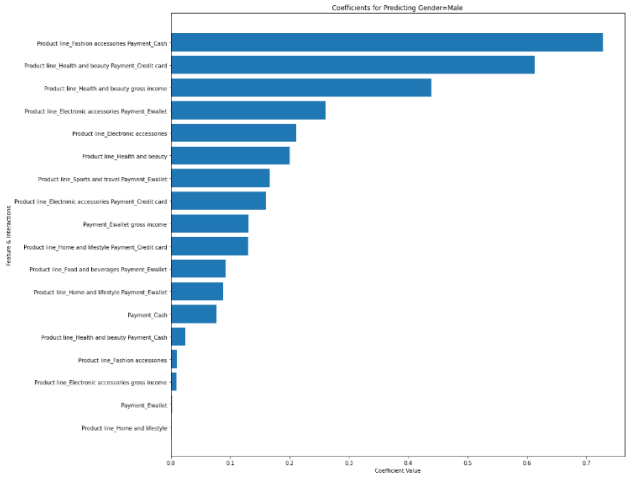

Figure 11. Bar graph showing the feature coefficients of the degree two interactions with Gender being ‘Male’.

From the plotted bar chart, we can see the magnitude of each coefficient. The attributes with the highest coefficients are the most informative when predicting gender (being Male). Interaction terms with significant coefficients indicate that the relationship between the two features has a considerable effect on predicting gender being Male. Hence, from the above plot, it is clear that the chances of Gender being Male are highest when the Product line is Fashion Accessories and Payment is Cash. It also seems like gross income has little influence on the gender being male.

4. When classifying the Customer type, attributes with large positive coefficients indicate a strong positive relationship with the "Normal" customer type.

   Attributes with large negative coefficients indicate a strong negative relationship with the "Normal" customer type. Features with coefficients furthest away from 0 (whether positive or negative) have the strongest impact on the prediction.

   From the plotted coefficients, it is apparent that Gender attribute is more impactful on the Customer type being Normal.

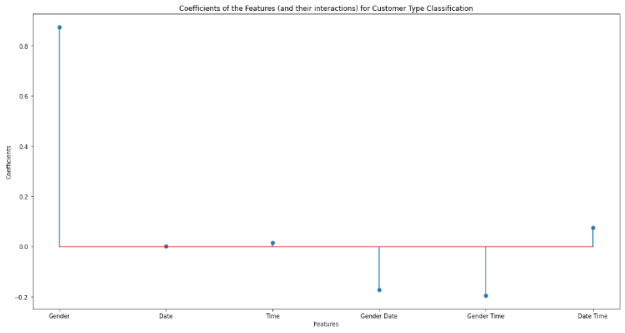

Figure 12. Stem plot showing the feature coefficients of the degree two interactions with Customer type being ‘Normal’

VI. Conclusions

From the observations made earlier, from the model results and interactions, when looking out for gross income, it is better to make decisions based on ‘Total’ and ‘cogs’ features as suggested by the Lasso Regularizer, with more importance on the ‘Total’ feature.

Predicting Unit Price can also be made easier by considering the ‘Total’ and ‘Quantity’ features. With ‘Total’ having a positive impact on the Unit prices and ‘Quantity’ having a negative impact on the Unit price. These conclusions suggest making business decisions accordingly. These conclusions are again suggested by the Lasso Regularizer meaning that these models can also be implemented in other business fields.

While understanding customer behavior, it is apparent that Males are more prone to have Fashion Accessories and pay with cash. So, if a customer buys a fashion accessory with cash, chances are that the customer is a male. With this information, even retailers can implement practices that enable male customers to buy these product lines easily with cash.

It is also apparent that Males are less likely to be members. So, it is suggestible to promote females to become members since they are more susceptible to being a member. These insights lend themselves to targeted marketing campaigns that can improve businesses.

References

[1] Supermarket Sales Dataset. (2022). Retail Chain Archive.