# Insights from Supermarket Sales Data: Regression and Classification Approaches

Vijay Abhinav Telukunta
Computer Science Engineering
*University of Florida*
Gainesville, USA
vtelukunta@ufl.edu

***Abstract*— This project explores supermarket sales data to gain insights and build predictive models. We preprocess the data, apply linear regression and logistic regression. We analyze the impact of various features on target variables like gross income, unit price, gender classification, customer type classification, day-of-purchase etc.**

## I. Introduction

This report focuses on a comprehensive analysis of a supermarket sales dataset. The primary objectives are data preprocessing, data visualization, regression, and classification.

## II. Data Preprocessing

The data preprocessing is crucial to the machine learning modelling. The following preprocessing steps were applied to transform and modify the data:

### *A. Categorizing Date and Time*

The 'Date' attribute was converted into 'Day of Week' to understand the impact of specific days on sales. The 'Time' attribute was divided into four segments: 'Morning', 'Afternoon', 'Evening', and 'Night'. This addition of the 'Time segment' feature assists in capturing the temporal patterns of sales.

### *B. Redundant Feature Removal*

Certain features were found to be redundant and so to simplify the dataset, we dropped some features.

The 'Invoice ID' feature is unique for each row, it acts as an identifier and doesn't provide any patterns or meaningful information for analysis. Similarly, the 'gross margin percentage' feature contains the same value (constant) for all rows in the dataset. As it doesn't provide any variation, it was removed.

### *C. Scaling and Encoding*

Numerical attributes such as 'Quantity,' 'Total', 'cogs' and 'Rating' were standardized using standard scaling so that all features would be on same scale. The attributes 'gross income' and 'Unit price' are not scaled for linear regression models because they are the target variables. Input features need to be scaled to predict accurate target. Hence, we scaled 'gross income' and 'Unit price' while training logistic regression models since at that time, these features serve as input. For categorical attributes, we applied one-hot encoding to features like 'Branch,' 'City', 'Product Line,' 'Payment,' 'Day of week,' and 'Time segment.'

The choice of one-hot encoding was made as it treated all these categories as equally important, preserving the individuality of each category.

Ordinal encoding was performed for the 'Customer Type' and 'Gender' features because they are target variables for classification tasks.

After the preprocessing steps, the dataset now contains a set of key features that are scaled, encoded and relevant to our analysis, including Branch, City, Customer Type, Gender, Product Line, Total, Quantity, cogs, Rating, Day of Week, and Time Segment.

## III. Linear Regression: Predicting Gross Income

In this section, Multiple linear regression was applied to predict gross income. The dataset was divided into a training set (80%) and a test set (20%) using train- test split. Two versions of the model were trained. One without regularization and another with Lasso regularization.

### *A. Impact of Features on Gross Income:*

In this model, several features were found to have a significant impact on gross income. The impact of the following features on gross income is discussed below:

*1) Quantity:* The weight or coefficient for quantity is significantly positive. It suggests that an increase in quantity for a particular product purchased by a customer has a substantial positive effect on gross income. As customers buy more items, the overall sales revenue (gross income) increases.

*2) Unit Price:* The positive coefficient for unit price indicates that higher product prices contribute positively to gross income. It implies that customers who purchased more expensive items generated higher gross income.

*3) Day of Week:* The coefficients for different days of the week suggest variations in customer spending patterns. For instance, customers who shop on Tuesday, Thursday and Saturday tend to contribute positively to gross income which means they purchase more on these days. In contrast, Sunday and Wednesday shoppers have a negative effect, possibly indicating lower spending on this day.

*4) Cash:* Customers paying through cash have a positive effect on gross income. It could indicate that cash transactions bypass transaction fees thereby increasing revenue compared to making payments using credit card.

*5) Time slot:* Different time slots have varying effects on gross income. Afternoon and evening shoppers contribute positively to gross income, which means they could be

spending more. While morning and night shoppers have a negative effect.

*6) Product line:* Different product lines have varying impacts on gross income. Home and lifestyle products have a positive effect, which means customer tends to buy more of these. Whereas certain categories like food and beverages, electronic accessories, sports and travel and health and beauty have negative effects.

### *B. Hyperparameter Tuning for Lasso Regularization*

Hyperparameter tuning for the best value of lambda was conducted using grid search. Lambda is the regularization hyperparameter which controls the strength of Lasso regularization.

Grid search explored a range of alpha values and identified the optimal value of lambda to be 0.087. The grid search model searches for lambda by 1000 even spaced integers in the range [0.05,0.2].

For model selection and evaluation, the coefficient of determination, $R^2$ (R-squared), was used as a key metric of success. The findings of the $R^2$ scores are presented in the TABLE I.

TABLE I. SCORES AND CI OF GROSS INCOME PREDICTOR

| Data | Model | R2 Score | 95% Confidence Interval |
|---|---|---|---|
| Train Set | Without Lasso | 0.8915989865026306 | (0.8684527077561283, 0.8944050981320147) |
| Train Set | With Lasso | 0.8896394278508686 | (0.8732725053022183, 0.8978864455877122) |
| Test Set | Without Lasso | 0.8913088363324457 | (0.8141377647513237, 0.8855238116286008) |
| Test Set | With Lasso | 0.892171966785714 | (0.8282937358129677, 0.8982480777484378) |

## IV. LINEAR REGRESSION: PREDICTING UNIT PRICE

In this section, Multiple linear regression was applied to predict unit price. The dataset was divided into a training set (80%) and a test set (20%) using train test split. Here, unit price is the target variable and remaining features form the feature matrix.

Two versions of the model were trained. One without regularization and another with Lasso regularization.

### *A. Impact of Features on Unit Price:*

In this model, several features were found to have a significant impact on unit price. The impact of the following features on unit price is discussed below:

*1) Gross Income:* The positive coefficient suggests that there is a positive relationship between the unit price and gross income. As gross income increases, the unit price tends to increase as well.

*2) Quantity*: The significant negative coefficient indicates an inverse relationship. As the quantity of products purchased increases, the unit price for that product decreases.

*3) Day of Week:* Unit prices tend to decrease on Tuesdays as it has negative effect with unit price.

*4) Cash:* There is a negative effect of cash on predicting unit price. It might imply that as more customers make payment through cash, inflow increases and unit price decreases.

### *B. Hyperparameter Tuning for Lasso Regularization*

Hyperparameter tuning was conducted using grid search technique. The primary hyperparameter under consideration was lambda which controls the strength of Lasso regularization. Grid search explored a range of alpha values by taking 500 evenly spaced integers from [0.01,0.5] and identified the optimal value of lambda to be 0.272.

For model selection and evaluation, the coefficient of determination, $R^2$ (R-squared), was used as a key metric of success. The findings of the $R^2$ scores and 95% confidence interval are presented in the TABLE II.

TABLE II. SCORES AND CI OF UNIT PRICE PREDICTOR

| Data | Model | R2 Score | 95% Confidence Interval |
|---|---|---|---|
| Train Set | Without Lasso | 0.7838946460322088 | (0.7400927434606275, 0.799723054494722) |
| Train Set | With Lasso | 0.7797665837469525 | (0.7469053186043333, 0.8058673132010952) |
| Test Set | Without Lasso | 0.7848505164456147 | (0.6844845503325383, 0.7970756896745368) |
| Test Set | With Lasso | 0.7879557586887668 | (0.7100104690324487, 0.8186926811842625) |

## V. LOGISTIC REGRESSION: GENDER CLASSIFICATION

In this section, logistic regression was used for classifying gender. To improve the predictive power, a polynomial feature with degree 2 and interaction_only as true is used. interaction_only=True means there would be a combination of features with degree 2 like Health and beauty+Tuesday, Female+Fashion accessories etc which could imply meaningful relationships. This increased the feature space to 562 features or attributes. Initially, we didn't scale the attributes: Unit price and Gross Income because they were target variables in linear regression. But now they are part of input feature matrix. So, we applied preprocessing step again to scale the Unit price and gross income. The model is trained using logistic regression to classify the target variable 'Gender'.

We create a pipeline which has Polynomial Features step and Logistic Regression step. After training the model, the following are the top 10 features which contributed mainly to the classification of gender. The features and their corresponding coefficients are shown in TABLE III. The corresponding plot of the coefficients are shown in Fig.1

Some features have positive coefficients, and some features have negative coefficients. Positive coefficients imply the probability of customer being classified as male. Negative coefficients mean decreasing likelihood of being classified as male which implies female. The coefficient implies the weight for that feature combination. Since till degree 2 is allowed so the feature combination would contain either one feature or two features maximum which impact in classifying gender target variable.

TABLE III. FEATURES AND COEFFICIENTS OF GENDER CLASSIFIER

| Feature | Coefficient |
|---|---|
| Home and lifestyle+Monday | -1.1549 |
| Fashion accessories+Saturday | -0.93651 |
| Electronic accessories+Tuesday | 0.87998 |
| Fashion accessories+Evening | 0.86064 |
| Electronic accessories+Wednesday | 0.83056 |
| Health and beauty + Friday | 0.82194 |
| Food and beverages+friday | -0.81103 |
| Sports and Travel+Friday | 0.76414 |
| Fashion accessories + Monday | 0.76301 |
| Electronic accessories+Thursday | -0.72692 |

Fig. 1. Plot of Feature vs Coefficient for the gender classifiers.

The significance of different features and their interaction as per the coefficients are given below

*1) Home and lifestyle + Monday:* Home and lifestyle purchases on Monday have negative effect of classifying gender as male.

*2) Fashion accessories + Saturday:* Fashion accessories purchase on Saturday has negative effect of classifying gender as male.

*3) Electronic accessories + Tuesday:* Electronic accessories purchased on Tuesday has positive effect on classifying gender as male. So, males purchase electronic accessories more on Tuesday.

*4) Fashion accessories + Evening:* Fashion accessories purchased on evenings have higher probability of classifying gender as male.

*5) Electronic accessories + Wednesday:* Electronic accessories purchased on Wednesday have higher probability that is purchased by male.

The Accuracy scores and its 95% confidence interval of training and test set is shown in TABLE IV.

TABLE IV. SCORES OF GENDER CLASSIFIER

| Data | Model | Accuracy Score | 95% ConfidenceInterval |
|---|---|---|---|
| Train Set | Logistic Regression | 0.76125 | (0.4530671657212499, 0.5444328342787501) |
| Test Set | Logistic Regression | 0.505 | (0.4773059442237334, 0.6526940557762665) |

## VI. LOGISTIC REGRESSION: CUSTOMER TYPE CLASSIFICATION

In this section, logistic regression was used for classifying customer type. To improve the predictive power a polynomial feature with degree 2 and interaction_only as true is used. This increased the feature space to 562 features.

Here, the target variable is Customer Type which has integer encoding either 0 or 1. The remaining features or attributes form the feature matrix.

After training logistic regression model to classify customer type, the following are the top 5 feature combinations which have high coefficients which impact customer type in positive or negative way.

TABLE V. FEATURES AND COEFFICIENTS OF CUSTOMER TYPE CLASSIFIER

| Feature | Coefficient |
|---|---|
| Fashion accessories+Monday | -1.108 |
| Health and Beauty + Friday | -1.080 |
| Sports and Travel + Wednesday | 1.0636 |
| Health and Beauty + Sunday | 1.0343 |
| Fashion accessories + Tuesday | 0.8888 |

Fig. 2. Plot of Feature vs Coefficient for the customer type classifiers.

The significance of different features and their interaction as per the coefficients are given below.

1) *Fashion accessories+Monday* : Fashion accessories purchased on Monday have negative effect on classifying customer as Normal.

2) *Health and Beauty + Friday*: Health and beauty products purchased likely by members on Friday.

3) *Sports and Travel + Wednesday:* Normal customers purchase Sports and Travel products on Wednesday.

4) *Health and beauty + Sunday:* Normal customers likely buy health and beauty products on Sunday.

5) *Fashion accessories + Tuesday:* Normal customers buy fashion accessories on Tuesday.

The Accuracy scores are calculated by comparing prediction values and target values for the data and its 95% confidence interval is also calculated. Results are shown in TABLE VI.

TABLE VI. SCORES OF CUSTOMER TYPE CLASSIFIER

| Data | Model | Accuracy Score | 95% Confidenceinterval |
|---|---|---|---|
| Train Set | Logistic Regression | 0.7425 | (0.43915313229233166, 0.5133468677076684) |
| Test Set | Logistic Regression | 0.525 | ((0.39385717996713043, 0.6161428200328696) |

```
# build random forest classifier

from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(random_state=53)
param_grid = {'n_estimators': np.arange(10,200,5)}

grid_search = GridSearchCV(rf_model, param_grid,
                           scoring='accuracy',
                           n_jobs=-1,
                           cv=10,
                           refit=True)
grid_search.fit(X_train, t_train)

grid_search.best_params_
```

## VII. CLASSIFICATION: DAY-OF-PURCHASE PREDICTION

In this section, a multi class classifier is trained to determine the day of week when the customer makes a purchase. This is done mainly by using 2 classifiers, Multi class logistic regression and random forest classifier.

The dataset was divided into training and testing sets using an 80:20 split ratio using train_test_split function.

### *A. Multi Class Logistic Regression:*

In The Logistic regression, made use of the "multi_class" parameter, set as "multinomial," to enable the prediction of multiple classes corresponding to different days of the week. The target variable here is Day of week. It was already encoded as one hot encoding. Hence, need to change it to integer encoding since now it is target variable. Used grid search strategy, exploring different hyperparameter combinations.. The key hyperparameters considered were the regularization term "C" and the penalty type, including "l1" and "l2." The hyper parameter values are shown in TABLE VII.

TABLE VII. HYPER PARAMETER VALUES FOR MULTI CLASS LOGISTIC REGRESSION

| Model | C | PENALTY |
|---|---|---|
| Grid Search Model | 0.3303030303030303 | L1 |

The model was trained on the training dataset to predict the day of purchase. Cross-validation was conducted to assess the model's generalization ability. Used solver=’saga’ for faster optimization and convergence.

The model performed poorly on the data obtaining very low accuracy score and it indicates model is underfitting.

The accuracy scores and its 95% confidence intervals for both the training and test set is shownin TABLE VIII.

TABLE VIII. SCORES OF MULTI CLASS LOGISTIC REGRESSION

| Data | Model | Accuracy Score | 95% Confidence Interval |
|---|---|---|---|
| Training Set | Logistic Regression | 0.22375 | (0.1376389128323543, 0.1773610871676457) |
| Test set | Logistic Regression | 0.155 | (0.08458869233302448, 0.19541130766697556) |

A random Forest Classifier is used to predict multiple classes corresponding to different days of the week. Employed a Grid Search strategy, exploring different hyperparameter combinations. The key hyperparameters considered were n_estimators. The hyper parameter values for n_estimator was 40 which best fits the model.

The model was trained on the training dataset to predict the day of purchase. Cross-validation was conducted to assess the model's generalization ability.

The model is performing extremely well on training data but has poor accuracy on test data which implies that model is clearly overfitting. The accuracy scores and its 95% confidence intervals for both the training and test set is shown in TABLE X.

TABLE X. SCORES OF RANDOM FOREST CLASSIFIERS

| Data | Model | Accuracy Score | 95% Confidence Interval |
|---|---|---|---|
| Training Set | Random Forest | 1.000 | (0.12761049873589952, 0.2098895012641005) |
| Test Set | Random Forest | 0.13 | [(0.11362458231259578, 0.2363754176874042) |