Analysis of Product's Sale Transactions Using Regression Models: Linear and Logistic Regression Models

1. Introduction: In the contemporary business landscape, accurate sales prediction is crucial for effective decision-making, resource allocation, and strategic planning. Machine learning models, particularly linear and logistic regression, are widely employed for sales prediction tasks due to their simplicity and interpretability. This report presents a comprehensive evaluation of linear and logistic regression models applied to a sales dataset loaded from the UCI Machine Learning Repository. The analysis includes data preprocessing, model training, evaluation metrics, and performance visualization.

2. Linear Regression Model and Logistic Regression: Linear Regression is a statistical model that predicts a continuous dependent variable based on one or more independent variables. It assumes that there is a linear relationship between the inputs (independent variables) and the output (dependent variable) (Shalev-Shwartz et al., 2014). Logistic Regression, unlike Linear Regression, is used for binary classification tasks. It predicts the probability that an observation belongs to one of two classes. If the probability is more significant than 0.5, the observation is classified into the positive class; otherwise, it is classified into the negative class (Shalev-Shwartz et al., 2014).

2.1 Input Data: The Products Sale Transactions dataset is downloaded from the UCI Machine Learning Repository. It includes weekly purchased quantities of 800 over products over 52 weeks. https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly.

2.2 Data Preprocessing: Data preprocessing involves splitting the dataset into features and target variables.

2.2.1 Features Selection: The features represent the number of weeks, we have 52 weeks in the dataset, and the target is the total sales of the products. This is followed by splitting the data into training and test sets to ensure the model can be evaluated on unseen data.

2.2.2 Normalization and Scaling: Feature scaling is performed to standardize the range of the data features, which helps in the convergence of the model. We observed that the data is being preprocessed by standardizing the features, which is a crucial step for both types of Regression, particularly when the features have different scales. This preprocessing can lead to more reliable estimates of coefficients and better model performance.

Handling missing values

2.3 Data Mining: We predicted the total sales of products purchased in 52 weeks using Linear Regression model to sum up weekly sales data to get a total sales figure for each product. Logistic Regression categorize these total sales into high (1) and low (0) based on a median threshold.

2.4 Post processing (Information Visualization): For the evaluation of linear regression model, we used Root mean squared error (RMSE) and R^2 matrix, and for logistic regression we used accuracy, precision, recall and F1 matrix. Table 1 reported the evaluation of regression models.

Table 1: Evaluation of Regression Models

Regression	**Matrix**	**Results (Scores)**
Linear Regression	Mean Squared Error (RMSE):	1.62e-13
Linear Regression	R^2	1.0
Logistic Regression	Accuracy	0.9693
Logistic Regression	Precision	0.9452
Logistic Regression	Recall	0.9857
Logistic Regression	F1	0.965

Figure 1 shows the scatter plot that illustrates a Linear Regression model's performance on sales data, with predictions closely aligned with actual sales. The proximity of data points to the diagonal line indicates accurate predictions. Minimal scatter signifies consistent model performance across various sales values. Overall, the model exhibits high predictive accuracy for total sales.

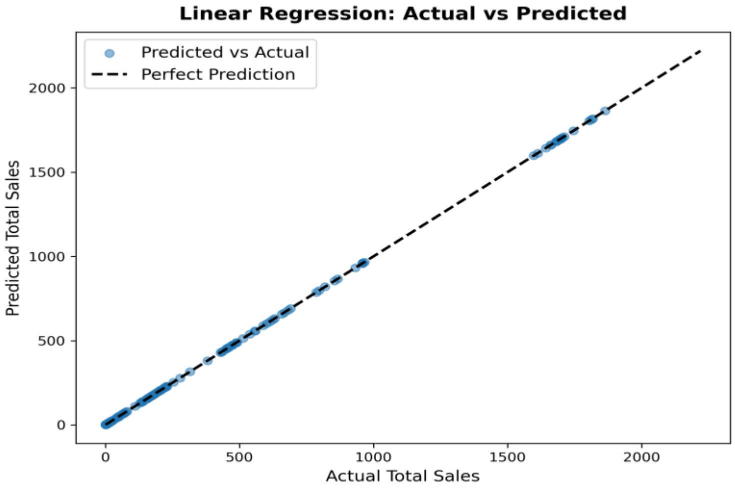

Figure 1: Predicted Total Sales using Linear Regression

Figure 2 shows the confusion matrix of Logistic Regression that shows a high number of true predictions (TN and TP) and a low number of false predictions (FN and FP), indicating good performance.

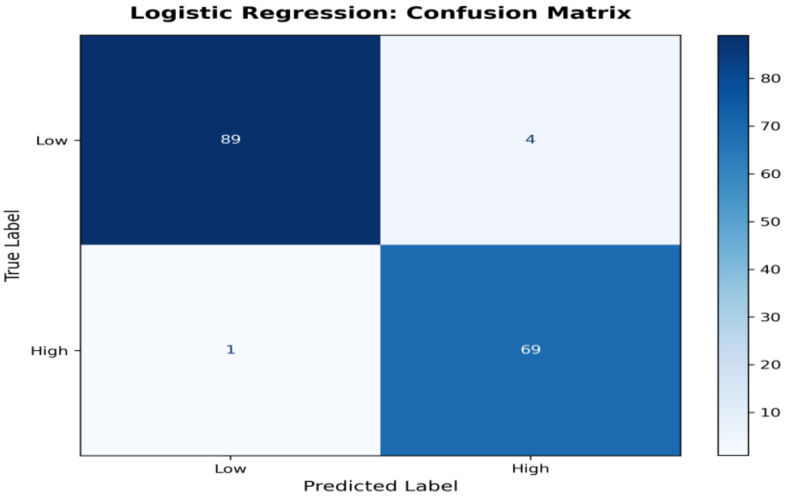

Figure 2: Performance of Logistic Regression in terms of confusion matrix

3. Conclusions: In conclusion, this report provided a comprehensive evaluation of linear and logistic regression models for sales prediction. Both models demonstrated reasonable performance, with linear regression predicting total sales and logistic regression classifying sales as high or low. The insights gained from this analysis can inform future modeling efforts and aid in making informed business decisions regarding sales forecasting.

4. Reference

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.