CSE422: Artificial Intelligence

Project Name: EDA Selling Price Prediction

Section: 09

Submitted By: Group 4

Name	ID
Faiza Binte Arif	21201498
Naeem Islam	21201171
Mohammad Omar Raihan	21141058
Shafin Islam Ohin	21241049

Submitted To:

Mr. Shayekh Bin Islam
Lecturer
Brac University

Mehran Hossain
Lecturer
Brac University

1. Introduction

In this project, "EDA Selling Price Prediction," we attempted to predict the selling price per unit of different products using an exploratory process. Accurate selling price forecasting is essential for many businesses, including manufacturing and retail, as it has a direct impact on inventory control, strategic planning, and revenue optimization. This study attempts to explore the depths of an extensive dataset using cutting-edge machine learning techniques in order to find patterns and relationships that can be used to accurately estimate an item's selling price. In addition to being a useful tool for companies looking to maximize their pricing tactics, selling price prediction is also a great way to apply data science and machine learning techniques to solve real-world problems.
The study aims to illustrate the capability of machine learning in extracting significant findings from diverse datasets and producing precise forecasts that can facilitate well-informed decision-making procedures by analyzing and modeling this data.

2. Dataset Description

The dataset used in this project is sourced from Kaggle and is titled "Sales from Different Stores." It can be accessed here
https://www.kaggle.com/datasets/kzmontage/sales-from-different-stores

Dataset Overview:

Number of Data Points: 99,457
Number of Features: 13
Task Type: The project focuses on a regression task, specifically predicting the selling price per unit of products.

Feature Details:

Quantitative Features:

- Age
- Quantity
- Selling Price per Unit
- Cost Price per Unit

Categorical Features:

- Invoice Number
- Invoice Date
- Customer ID
- Gender

- Category
- Payment Method
- Region
- State
- Shopping Mall

Correlation Analysys:

In the exploratory data analysis phase, a thorough examination of the correlation between various features was conducted, with some notable findings:

Strong Correlation between Cost Price and Selling Price: A perfect correlation score of 1.00 was observed between 'cost_price_per_unit' and 'selling_price_per_unit'. This suggests a direct relationship where the selling price is heavily influenced by the cost price, a common scenario in retail and sales industries.

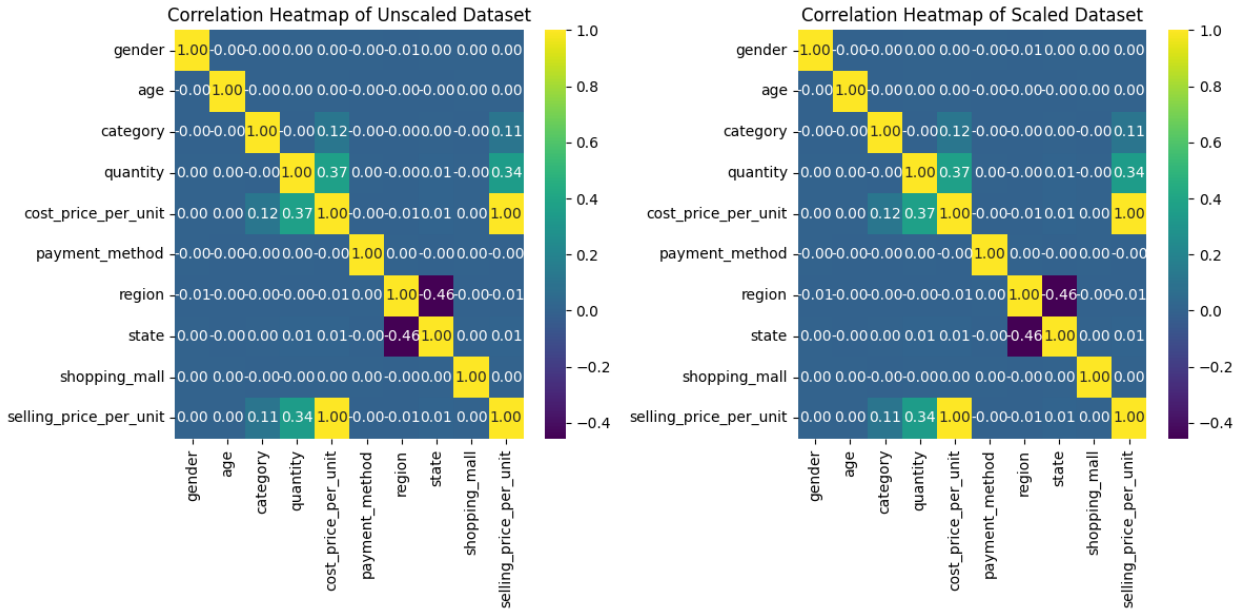
Moderate Correlation between State and Region: A correlation score of 0.46 was found between 'state' and 'region', indicating a moderate association. This is understandable as states are often grouped into regions based on geographical, cultural, or economic factors.

Correlation between Cost Price and Quantity: A correlation score of 0.37 was noted between 'cost_price_per_unit' and 'quantity'. This relationship could be indicative of bulk purchasing patterns or economies of scale in pricing.

Correlation between Selling Price and Quantity: There was also a correlation of 0.34 between 'selling_price_per_unit' and 'quantity', which could reflect pricing strategies based on quantity sold.

Other Correlations: No other significant correlations were found in the dataset.

These findings, particularly the strong correlation between cost and selling prices, are logical and align well with real-world scenarios. A heatmap was utilized to visually represent these correlations, offering an intuitive understanding of these relationships and their strengths.



3. Dataset Preprocessing

Several steps were taken during the preprocessing phase to prepare the dataset for effective machine learning modeling. These steps are essential for ensuring data quality and usability, as well as improving model performance. The following preprocessing procedures were carried out:

Rearranging the Target Column:

The target column 'selling_price_per_unit' was moved to the end of the dataframe for convenience and clarity.

Dropping Irrelevant Columns:

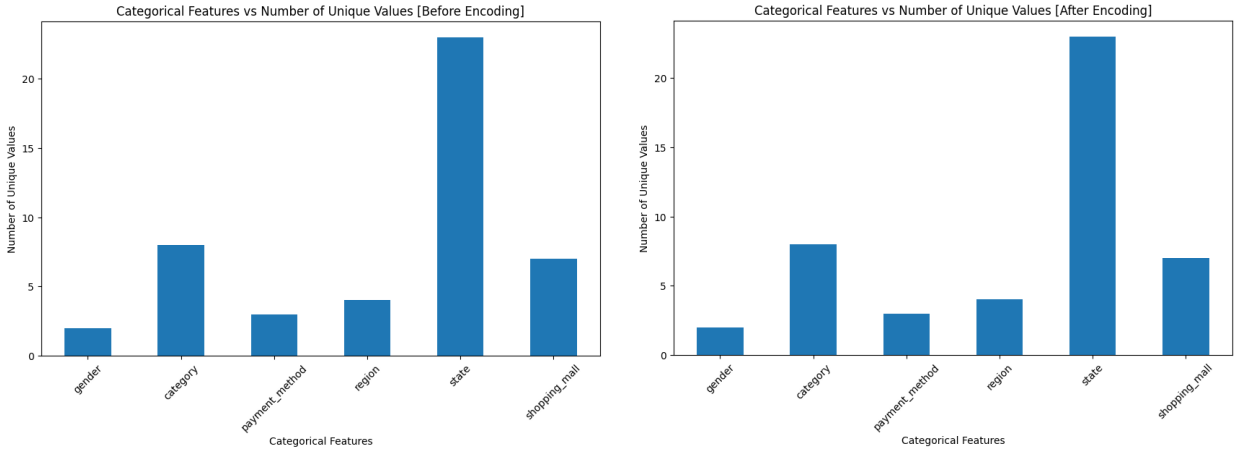
Columns that were deemed irrelevant for the prediction task, such as 'invoice_no', 'invoice_date', and 'customer_id', were dropped from the dataset.

Handling Null Values:

It was determined that the dataset did not contain any null values, hence no imputation or removal of null values was required.

Encoding Categorical Variables:

The dataset contained several categorical columns: 'gender', 'category', 'payment_method', 'region', 'state', and 'shopping_mall'. These were encoded to numerical values using Label Encoding, as most machine learning algorithms require numerical input.



These preprocessing steps were essential in transforming the raw data into a format suitable for machine learning algorithms, ensuring the data's integrity and relevance to the predictive modeling task.

4. Feature Scaling

Feature scaling is an important step in many machine learning pipelines, especially when using algorithms that are sensitive to input data scale. Feature scaling was used in this project to standardize the range of independent variables or features in the dataset. The following are the steps and rationale for this process:

Standardization Using StandardScaler:

The StandardScaler from scikit-learn was employed. This scaler standardizes features by removing the mean and scaling to unit variance. Standardization was necessary as it normalizes the data and ensures that each feature contributes equally to the model, which is especially important for models like Support Vector Machines and Neural Networks.

Application of Scaling:

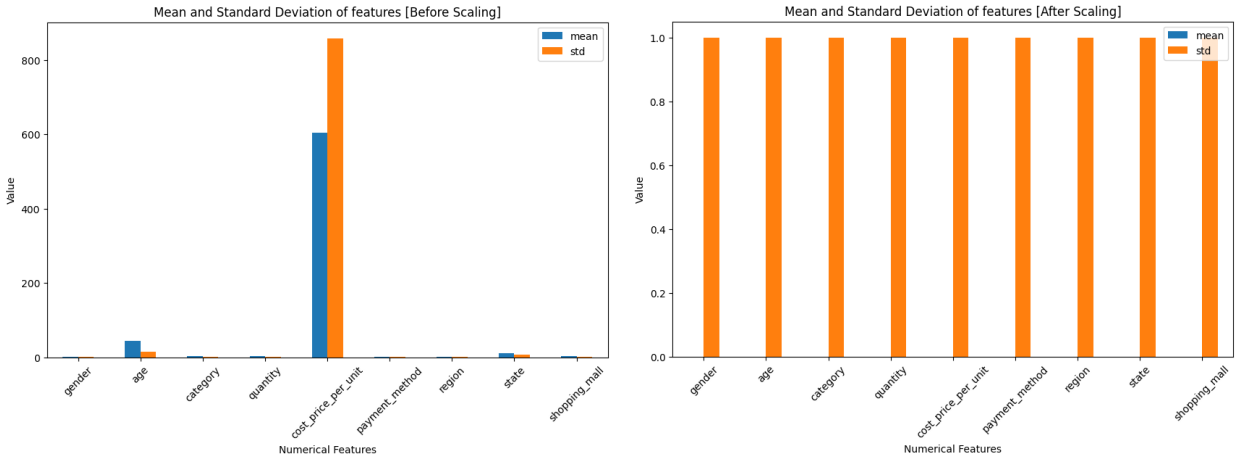
The scaling was applied to all features except the target variable 'selling_price_per_unit'.

Reforming the DataFrame:

A new DataFrame final_df was created with the scaled features. The target variable was appended to this DataFrame without scaling, as it is the dependent variable we aim to predict.

Visualizing the Scaling Effect:

To illustrate the impact of feature scaling, plots depicting the mean and standard deviation of features before and after scaling is included below. These visualizations will clearly show how the data has been transformed to a standard scale, enhancing the model's ability to learn from the features effectively.



Feature scaling was a pivotal step in ensuring that the predictive models performed optimally and that all features were given equal importance in the learning process.

5. Dataset Splitting

The splitting of the data into training and testing sets is an important step in preparing the dataset for machine learning modeling. This split enables the model to be evaluated on unseen data, providing an estimate of its performance on real-world data. The following are the steps and details for splitting the dataset in this project:

Defining Features and Target Variable:

The features (X) and the target variable (y) were defined. 'X' includes all the columns except for the target variable 'selling_price_per_unit', while 'y' is solely the 'selling_price_per_unit'.

Splitting Ratio:

The dataset was split into training and testing sets with a 70-30 ratio, where 70% of the data was used for training the models, and the remaining 30% was set aside for testing. A 70-30 split is commonly used in machine learning as it provides a substantial amount of data for learning while retaining enough data to test and validate the models effectively.

Random Split with Fixed Random State:

The train_test_split function from scikit-learn was used for this purpose. A random_state was set to ensure reproducibility of the results.

Result of the Split:

The training set consisted of 69,619 samples (features and target), and the testing set comprised 29,838 samples.

This methodical approach to dataset splitting ensures that the models are trained on a diverse set of data points and are tested on different samples to gauge their generalization ability effectively.

6. Model Training and Evaluation:

The project included training and testing three distinct machine learning models: linear regression, decision trees, and neural networks. Each model was chosen for its distinct characteristics and ability to address regression issues. The following explains the training and evaluation process for each model:

Linear Regression:

This model is known for its simplicity and interpretability. It was trained on the dataset to establish a baseline for performance.

Decision Tree Regressor:

Decision trees are useful for their ability to capture non-linear relationships. The model was trained to understand how it performs in comparison to the linear model.

Neural Network (MLPRegressor):

A more complex model, the Neural Network, specifically a Multi-layer Perceptron regressor, was chosen for its ability to model highly complex relationships in the data.
The model was allowed a significant number of iterations (500) for training to ensure adequate learning.

Evaluation Process:

Each model was trained on the training set and then used to make predictions on the test set.
The evaluation focused on both the error rate (MSE) and the proportion of variance explained (R2), providing a balanced view of model performance.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

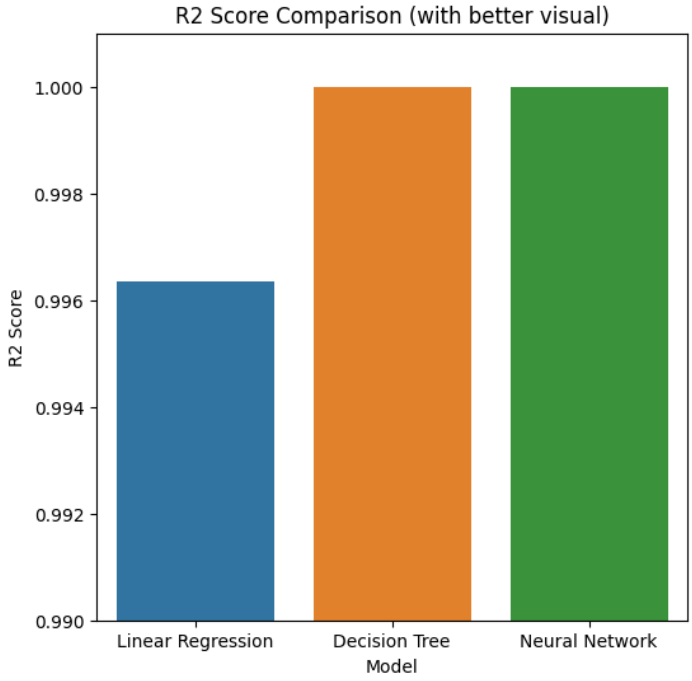
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This comprehensive training and evaluation process allowed for an in-depth comparison of how each model performs in predicting the selling price per unit and understanding the strengths and weaknesses of each approach in the context of the given dataset.

7. Model Selection and Comparison Analysis

The project involved a thorough comparison and analysis of three machine learning models: Linear Regression, Decision Tree, and Neural Network. The initial performance of these models was evaluated based on two key metrics: Mean Squared Error (MSE) and R-squared (R2).

The results were as follows:
Linear Regression: MSE of 3246.854222 and an R2 Score of 0.996351.
Decision Tree: MSE of 0.055407 and an R2 Score of 1.000000.
Neural Network: MSE of 4.041713 and an R2 Score of 0.999995.



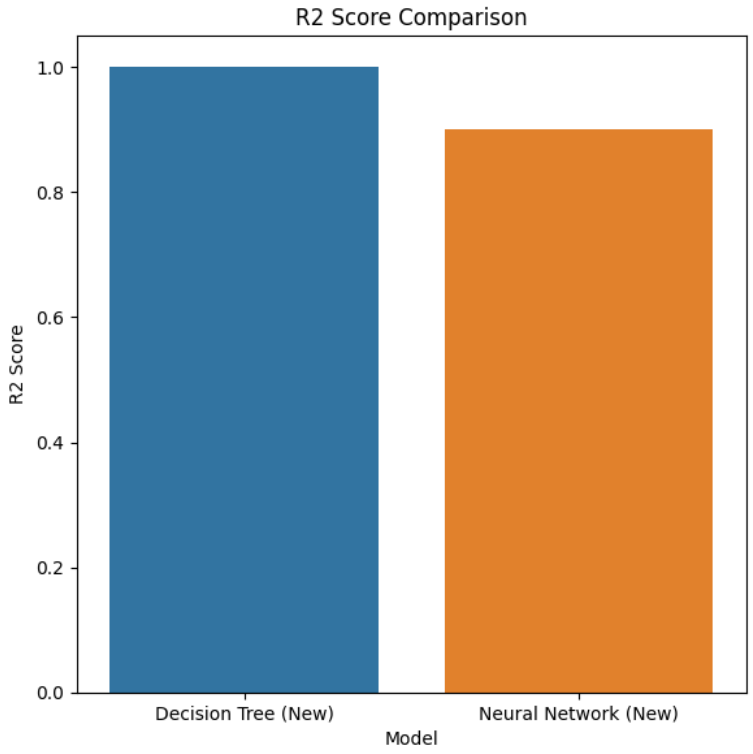
Based on these metrics, the Decision Tree model significantly outperformed the other two, exhibiting nearly perfect predictive accuracy. The Neural Network also demonstrated excellent performance, though not as outstanding as the Decision Tree.

Impact of Feature 'cost_price_per_unit':

A crucial observation was the high correlation between 'cost_price_per_unit' and 'selling_price_per_unit'. To assess the impact of this feature, the Decision Tree and Neural Network models were retrained after excluding 'cost_price_per_unit' from the dataset. The performance of the models after this modification was as follows:

Decision Tree:
MSE: 0.08474188618540118
R2 Score: 0.9999999047635091

Neural Network:
MSE: 88546.12337963954
R2 Score: 0.9004881475785862



The Decision Tree model continued to perform exceptionally well, whereas the performance of the Neural Network model dropped notably. This variation in performance indicated that the Neural Network model was previously over-relying on the 'cost_price_per_unit' feature, leading to its reduced effectiveness when this dominant feature was removed.

Outcome from Model Comparison:

Decision Tree as the Best Model: The Decision Tree model proved to be the most robust and consistent performer, both before and after the removal of the 'cost_price_per_unit' feature. Its ability to maintain high accuracy indicates its reliability and suitability for this prediction task.

Neural Network's Dependency: The drop in performance of the Neural Network model suggests a dependency on the 'cost_price_per_unit' feature, highlighting the importance of feature selection and the risk of model bias towards dominant features.
Overall, this comparison and analysis shed light on the strengths and weaknesses of each model and emphasized the significance of careful feature selection in predictive modeling.

8. Conclusion

This project, which centered on predicting product selling prices per unit, successfully navigated through the various stages of a machine learning project, from data preprocessing to model training and evaluation. The journey resulted in insightful findings that not only shed light on the capabilities of various machine learning models, but also highlighted the importance of careful feature selection and model evaluation.

Key Findings:

- The Decision Tree model emerged as the most robust and consistent performer in our project. Its ability to adapt to changes in the feature set and maintain high accuracy underlines its effectiveness for this specific predictive task.
- The performance of the Neural Network model, although impressive initially, was notably affected by the removal of the 'cost_price_per_unit' feature. This highlighted the model's sensitivity to certain dominant features and underscored the need for balanced feature representation in training data.
- The correlation analysis, particularly the strong link between 'cost_price_per_unit' and 'selling_price_per_unit', played a pivotal role in understanding the dynamics of the dataset and guiding the feature selection process.

Finally, this project not only met its goal of predicting product selling prices, but it also provided a thorough understanding of the complexities involved in machine learning projects, from data preparation to model evaluation and selection. These discoveries pave the way for further analysis and continuous improvement in the field of predictive modeling.