Supervised Learning and Ensemble Techniques

Week 1 Practice Project (Part-1)

Topics Covered:

- **Linear Regression**

Domain:

E-commerce

Objective:

Predicting the price of a certain product with the help of a linear regression model on an E-commerce dataset.

Problem Statement:

The data contains the orders made in olist (i.e. https://olist.com/). The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions(i.e.price, payment and freight performance, payment_type, product_id, order_id, payment_sequential, shipping_limit_date, seller_id, payment_installment, payment_values). Olist wants to predict the price of new items that are coming to its inventory.

Feature Details:

1. order_id: order unique identifier
2. order_item_id: sequential number identifying number of items included in the same order.
3. product_id: product unique identifier
4. seller_id: seller unique identifier
5. shipping_limit_date:Shows the seller shipping limit date for handling the order over to the logistic partner.
6. price: item price(Target)
7. freight_value: item freight value item
8. payment_sequential: a customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.
9. payment_type:method of payment chosen by the customer.

10. payment_installments: number of installments chosen by the customer.

Tasks to be performed:

- **Data Loading and Exploration.**
    1. Import all the necessary libraries.
    2. Load the CSV file (i.e. Retail.csv) and display the first 5 rows of the dataframe.
    3. Check the info of the data frame.
    4. Check the shape of the data frame. Check if there are any missing values in any column of the dataset.
    5. Check the statistical summary of the data frame and write your findings.
    6. Plot the count plot of the ‘payment_type’. Which payment mode is used the lowest number of times to purchase the products.
    7. Do a bivariate analysis between the ‘payment_type’ column and ‘payment_installments’ Column. Plot a bar plot that represents all the payment_type labels of all “payment_installments”.
    8. Plot a pair plot of all the columns using hue as the ‘payment_type’ and share your insights.
    9. Plot a heatmap that consists of the correlation between all the columns of the data frame.
    10. Do frequency encoding of the column”payment_type” and store it in a new column name as “payment_freq”.

        (i.e. Frequency encoding is an encoding technique that takes frequency distribution into account. (i.e. suppose we have a data frame that has L rows and a column c1 of the same data frame consists of 2 labels P & Q. Now if P has occurred M times and Q has occurred N times then its frequency encoding will be M/L and the frequency encoding of the Q will be N/L respectively.)
- **Model Building and Evaluation.**
    1. Drop(**'order_id','order_item_id','product_id','seller_id','shipping_limit_date','payment_type'**) columns from the dataset.
    2. Store the target column(i.e. price) in the y variable and the rest of the columns in the X variable.
    3. Split the dataset into two parts(i.e. 70% train and 30% test) using random_state=42.
    4. Train a linear regression model and print the r2_score for both the train and test set.