Programming for Data science

Project Guild-line: Prediction of Books Price
Group 2

Project overview	The goal of this project is to study how prices of books are defined based on which feature, and also help publishers and authors make informed decisions about pricing their books to maximize sales and profitability while remaining competitive in the market..
Project objective	The objective of this project is to create a book price prediction model with high accuracy. The model should take relevant book features as input and output the predicted price.
Data collection	We will be using the "BookShopDataset" which we have scrapped from Bookshop website for this project. The dataset contains information on 2393 books, including features like title, Page, Format, Language, publication Date, Publisher, Author, Dimensions, Categories and the corresponding book prices. The dataset has already been preprocessed and is available in CSV format.

Data Preprocessing	We have performed the following data preprocessing steps:
	– Check missing value and Extract Length, Width, Height from Dimensions feature
	– Cleaning Categories column
	  – Replace triple space between each category type by commas and remove the last extra one at the end
	  – Remove unnecessary bracket text from categories
	  – Remove sub section (located after " - ")
	– Remove outlier in both categories variables and numerical

Data Exploratory	– Feature Engineering
	  – Change the format of publish date into day count since published.
	  – Binning the Surface Area into Cover size type and Height into Thickness Type
	  – We do ordinal encoding to covert cover size to three class which Small, Medium, Large and same to thickness feature to Thin, Medium, Thick.
	  – Target encoding on book categories.
	– Features Visualization
	  – we want to observe the popularity format of book over the year whitin hardcover, paperback, paperbounce.

Model Selection	– we have Selected appropriate machine learning algorithms, such as Multiple linear regression, random forests, or gradient boosting regression, support vector regression, Neural Network regression, Ridge and Lasso Regression for book price prediction. – Train the chosen models using the training set, optimizing hyperparameters using techniques like grid search – Split the dataset into training, validation, and testing sets using techniques like cross-validation to evaluate model performance effectively.
Model Evaluation	– Evaluate the trained models using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or mean absolute error (MAE). – Compare the performance of different models and select the best-performing model. – Analyze the strengths and weaknesses of the chosen model, considering factors such as interpretability, accuracy, and computational efficiency.