Computer Science Department
CS667 – Practical Data Science (CRN: 72872)
Fall 2025

Project #2 / Due 21-Oct-2025

Next step — transitioning from EDA to predictive modeling is a great applied exercise in your Practical Data Science course. Below is a complete **project writeup** for the **Machine Learning (Regression Analysis)** phase using the same **Retail Sales Dataset**.

Project Overview

In this phase of your project, you will extend your previous (Project #1) Exploratory Data Analysis (EDA) on the **Retail Sales Dataset** by applying **Machine Learning techniques** to build a **predictive regression model**.

The goal is to use the dataset as **training data** to predict a continuous numerical target variable — for example, **Total Sales Amount** (Revenue) based on various independent features.

You will **design**, **train**, and **evaluate** a regression model using modern **ensemble learning methods**, such as **XGBoost**, **Random Forest**, or **Gradient Boosting Regressor**.

Learning Objectives

By the end of this project, you should be able to:

- Prepare real-world retail data for machine learning.
- Perform **feature engineering** and **encoding** on categorical variables.
- Split the dataset into **training and testing sets**.
- Train and tune **ensemble regression models** (XGBoost, Random Forest, etc.).
- Evaluate model performance using regression metrics.
- Interpret feature importance and explain model predictions.

Project Steps

1. Data Preparation (This is mostly done during Project #1, just review it…)

- Load the dataset and inspect the structure (columns, datatypes, missing values).
- Remove duplicates or irrelevant columns.
- Handle missing data (e.g., impute with mean/median or drop).
- Convert currency or string values (e.g., “$123.45”) into numeric form.
- Convert categorical features (e.g., region, gender, product category) into machine-readable form using:
  - **Label Encoding** or
  - **One-Hot Encoding (pd.get_dummies or sklearn.preprocessing.OneHotEncoder)**.

- Create new derived features if meaningful (e.g., total items, day of week, weekend flag, season, etc.).

2. Feature Selection and Data Splitting

- Identify independent (predictor) variables and your target (dependent) variable.
- Split your data into **training** and **testing** sets, e.g.:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- Perform **feature scaling** (e.g., StandardScaler or MinMaxScaler) if needed.

3. Model Building

You will build and compare multiple regression models, focusing on **base** and **ensemble learning methods**:

Base Models (Simple) – Select One Model

- Linear Regression
- Decision Tree Regressor

Ensemble Models (Main Focus) – Select One Model

- **Random Forest Regressor**
- **XGBoost Regressor**
- **Gradient Boosting Regressor**

Train each model using the training data and evaluate their performance on the test data.

4. Model Evaluation

Evaluate models using regression performance metrics:

- **R^2 (Coefficient of Determination)**
- **MAE (Mean Absolute Error)**
- **MSE (Mean Squared Error)**
- **RMSE (Root Mean Squared Error)**

Visualize and compare model performance using:

- Actual vs. Predicted plots
- Residual plots
- Feature importance bar chart (especially for XGBoost or Random Forest)

5. Model Interpretation

- Interpret top predictive features (e.g., which features most influence total sales).
- Discuss the possible business meaning of these findings.

6. Reporting and Deliverables

Deliverables

Component	Description
Notebook/Script	Complete, well-commented notebook showing data preparation, model training, tuning, and evaluation.
Performance Summary	Table comparing model metrics (R^2, MAE, RMSE, etc.) for all models tested.
Visualizations	- Correlation heatmap - Feature importance plot - Actual vs. Predicted plot - Residuals plot
Written Report (1-2 pages)	Summarize your approach, data processing steps, chosen model, results, and business insights.
Optional Presentation	Short slide deck summarizing results for a business audience.

Key Takeaway

This project aims to help you transition from exploratory data analysis to predictive modeling — developing practical skills in **machine learning pipelines, feature engineering, and ensemble model interpretation**.

Dataset's (retail_sales.xlsx) Metadata Info

Transaction ID: A unique identifier for each transaction, allowing tracking and reference.
Date: The date when the transaction occurred, providing insights into sales trends over time.
Customer ID: A unique identifier for each customer, enabling customer-centric analysis.
Gender: The gender of the customer (Male/Female), offering insights into gender-based purchasing patterns.
Age: The age of the customer, facilitating segmentation and exploration of age-related influences.
Product Category: The category of the purchased product (e.g., Electronics, Clothing, Beauty), helping understand product preferences.
Quantity: The number of units of the product purchased, contributing to insights on purchase volumes.
Price per Unit: The price of one unit of the product, aiding in calculations related to total spending.
Total Amount: The total monetary value of the transaction, showcasing the financial impact of each purchase.