

Project Proposal: E-commerce Sales Prediction

1. Problem Statement

Accurate sales forecasting is critical for e-commerce businesses to make informed decisions related to inventory planning, marketing budget allocation, pricing strategies, and demand management. Inaccurate forecasts can lead to overstocking, stockouts, inefficient marketing spend, and lost revenue.

The goal of this project is to build a machine learning model that predicts **Units Sold** for e-commerce products based on historical sales, pricing, discounts, customer segments, and marketing spend. This is framed as a **supervised regression problem**, where the target variable is continuous.

2. Business Objective

The primary business objective is to **minimize prediction error in sales forecasts** so that stakeholders can better estimate expected demand.

Key stakeholders:

- Marketing team (campaign planning and ROI estimation)
- Operations & inventory planners
- Business analysts

Business question:

Given product, pricing, discount, customer segment, and marketing information, how many units are likely to be sold?

3. Dataset

The dataset used for this project is publicly available on Kaggle:

E-commerce Sales Prediction Dataset

<https://www.kaggle.com/datasets/nevildhinoja/e-commerce-sales-prediction-dataset>

Key features include:

- **Date** – date of the transaction
- **Product_Category** – product category
- **Price** – product price
- **Discount** – discount applied
- **Customer_Segment** – type of customer
- **Marketing_Spend** – marketing expenditure
- **Units_Sold (target)** – number of units sold

The dataset contains approximately 1,000 observations.

4. Success Metrics

The primary evaluation metric for this project is **cross-validated Mean Absolute Error (MAE)**.

Rationale:

- MAE is easy to interpret in business terms (average error in units sold)
- It is less sensitive to outliers than RMSE
- Cross-validation provides a more robust estimate of generalization performance than a single train–test split

Secondary metrics such as RMSE and R² were explored but not emphasized, as MAE was selected as the primary comparison metric across all models.

5. Methodology

The project follows the standard data science lifecycle:

5.1 Data Wrangling & Feature Engineering

- Parsing and cleaning date columns
- Extracting time-based features (month, day, weekday)
- Handling missing values
- Encoding categorical variables using one-hot encoding
- Scaling numerical features where appropriate

5.2 Exploratory Data Analysis (EDA)

- Distribution analysis of target variable (Units Sold)
- Relationship between discounts, marketing spend, and sales
- Category-level sales comparisons
- Identification of potential outliers

5.3 Modeling

The following regression models will be evaluated:

- **Baseline model** (mean predictor)
- **Linear Regression**
- **Random Forest Regressor**

Hyperparameter tuning is performed using **GridSearchCV**, and models are compared using **cross-validated MAE**.

6. Model Evaluation Strategy

- K-fold cross-validation is used to compute mean and standard deviation of MAE
- All models are evaluated using the same metric for fair comparison
- A held-out test set is used for sanity checks and final validation

The final model is selected based on the lowest cross-validated MAE, balanced with model complexity and interpretability.

7. Expected Outcomes

- A regression model capable of predicting units sold with reasonable accuracy
- Identification of key drivers of sales (e.g., price, discount, marketing spend)
- Actionable insights that can inform pricing and marketing strategies

8. Limitations & Future Work

- The dataset is relatively small, which may limit model generalization
- External factors such as seasonality, competition, and macroeconomic trends are not included

Future work could include:

- Incorporating additional external data sources
- Time-series-based forecasting approaches
- Advanced models such as Gradient Boosting or neural networks

9. Deliverables

- Cleaned and well-documented Jupyter notebooks
- Model metrics summary table
- Final written report summarizing findings and recommendations
- GitHub repository with reproducible code