Sales Prediction Using Machine Learning

1. Problem Understanding

The objective of this project is to build a machine learning model that predicts product sales using historical retail data. Accurate sales prediction plays a crucial role in retail businesses, as it supports inventory planning, demand forecasting, and informed business decision-making.

The dataset used in this project is a publicly available real-world retail dataset containing both numerical and categorical features related to products and outlet characteristics. The dataset also includes missing values and mixed data types, which reflect common challenges encountered in practical, real-world data scenarios.

This problem is formulated as a supervised regression task, where the objective is to predict a continuous target variable, Item_Outlet_Sales, based on multiple input features. The focus of this project is not solely on achieving high prediction accuracy, but also on demonstrating effective data preprocessing, appropriate model selection, robust evaluation, and clear interpretation of results.

2. Model Pipeline Description

Data Loading and Exploration

The dataset was loaded and initially explored to understand its structure, feature types, and the presence of missing values. Exploratory Data Analysis (EDA) was performed to examine relationships between key features and the target variable, such as the relationship between item price and sales, as well as the impact of different outlet types on sales performance.

Data Preprocessing

Missing values in numerical features were handled using mean imputation, while missing values in categorical features were filled using the mode. Inconsistent categorical values were standardized to ensure data uniformity. Categorical variables were transformed into numerical format using one-hot encoding. The target variable was explicitly separated from the input features to prevent data leakage during model training.

Train–Test Split

The dataset was divided into training and testing sets using an 80–20 split. This approach enables model evaluation on unseen data and provides an unbiased estimate of the model's generalization performance.

Model Selection and Training

Two machine learning models were trained and evaluated:

- Linear Regression was used as a baseline model due to its simplicity and ease of interpretation.
- Random Forest Regressor was used as the final model to capture complex, non-linear relationships between input features and sales.

3. Results & Evaluation Metrics

Model performance was evaluated using the following regression metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Model	MAE	RMSE	R^2 Score
Linear Regression	~944	~1274	~0.40
Random Forest	~760	~1094	~0.56

The Random Forest model outperformed Linear Regression across all metrics, indicating its ability to better model complex relationships present in the data.

4. Inference & Model Explanation

Feature importance analysis from the Random Forest model indicates that item price (MRP) and outlet-related attributes are the most influential factors in predicting sales. This observation aligns with real-world retail behavior, where pricing strategies and store characteristics play a significant role in shaping customer purchasing decisions.

Although the model demonstrates reasonable predictive performance, it does not incorporate temporal factors such as seasonal demand variations or promotional events. Future improvements could involve integrating time-based features or adopting time-series modeling techniques to further enhance prediction accuracy.

5. Conclusion

This project successfully demonstrates an end-to-end machine learning pipeline for sales prediction using real-world retail data. The results highlight the importance of effective data preprocessing and appropriate model selection when working with practical datasets. The Random Forest model achieved improved predictive performance compared to the baseline model while maintaining interpretability through feature importance analysis.