# MACS Sales Quantity Prediction using Machine Learning

## 1. Introduction & Problem Statement

Accurate sales quantity forecasting is critical for effective inventory management, pricing strategies, and marketing decisions in retail. This project aimed to develop a robust machine learning model to predict the sales quantity of various products across different stores using a comprehensive dataset containing product, customer, store, and environmental features.

---

## 2. Approach

**Data Cleaning & Preparation:** - Checked and imputed missing values using median imputation. - Ensured consistent data types across numerical and categorical columns.

**Exploratory Data Analysis (EDA):** - Visualized missing values heatmaps to confirm imputation. - Analyzed sales quantity distribution, revealing skewness requiring robust model handling. - Generated correlation heatmaps to understand feature relationships.

**Feature Engineering:** - Created derived features: `price_diff`, `discount_ratio`, `footfall_per_staff`, `weekend_footfall`. - These enhanced the dataset by capturing business-relevant signals.

**Model Selection:** - Chose LightGBM due to its efficiency and native handling of categorical variables.

**Hyperparameter Tuning:** - Applied Optuna for Bayesian optimization, improving the model's RMSE.

---

## 3. Feature Analysis & Insights

Using SHAP for interpretability: - Top impactful features were `revenue`, `actual_price`, `base_price`, `customer_income`, and `customer_footfall`. - Price and revenue features were most predictive, confirming domain expectations. - Customer income and footfall also contributed significantly.

---

## 4. Model Performance & Evaluation

The tuned LightGBM model achieved: - **RMSE:** 2.8301 - **MAE:** 0.4313 - **$R^2$ Score:** 0.9540

This indicates high predictive accuracy, with the model explaining ~95% of variance in sales quantity.

---

## 5. Final Model Summary

Final LightGBM model hyperparameters: - `learning_rate`: 0.0816 - `max_depth`: 5 - `num_leaves`: 32 - `colsample_bytree`: 0.8371 - `subsample`: 0.9128 - `reg_alpha`: 0.6319 - `reg_lambda`: 0.7907

The model was saved for deployment and used to generate test predictions saved as `aadya_result.csv`.

---

## 6. Visualizations

- Missing values heatmap
- Sales quantity distribution histogram
- Correlation heatmap
- Pairplots for key predictors
- SHAP feature importance plots

These visualizations validated data quality, highlighted key relationships, and ensured model interpretability.

---

## 7. Challenges Faced & Learnings

- Managing missing values and large categorical variables.
- Handling package compatibility issues with NumPy and SHAP.
- Using Optuna effectively to enhance model performance.
- Balancing accuracy with model complexity to avoid overfitting.

---

## 8. Conclusion & Future Work

A high-performing LightGBM regression model was successfully developed for sales quantity prediction, supporting inventory planning, discounting, and marketing strategy.

Future improvements could include: - Adding seasonality and granular holiday features. - Testing ensemble methods for further accuracy improvement. - Creating a retraining pipeline for continuous learning on new data.

---

## Attachments

- `aadya_result.csv` (final predictions)
- `aadya_MACS_SalesPrediction.ipynb` (full structured notebook)
- This report for PDF submission to MACS.

---

**Prepared By:** Aadya Jha