Sales Forecasting System: Project Documentation

1. Project Overview

This project aims to accurately forecast monthly sales at the item–outlet level for Big Mart, using historical sales data enriched with product and outlet attributes. Robust, actionable forecasts help drive inventory planning, pricing strategies, and marketing decisions.

2. Problem Definition

- **Goal:** Predict the variable `Item_Outlet_Sales` for a given set of item and outlet combinations.
- **Business Motivation:**
  - Reduce stockouts/overstock
  - Maximize revenue with better marketing and inventory control
  - Provide data-driven recommendations for resource allocation

3. Dataset Used

- **Source:** Kaggle Big Mart Sales Dataset
- **Files:** `Train.csv` (used for analysis/modeling), `Test.csv` (used for submission/prediction)
- **Key Columns:**
  - Product: `Item_Identifier`, `Item_Weight`, `Item_Fat_Content`, `Item_Type`, `Item_MRP`
  - Outlet: `Outlet_Identifier`, `Outlet_Establishment_Year`, `Outlet_Size`, `Outlet_Location_Type`, `Outlet_Type`
  - **Target:** `Item_Outlet_Sales` (only present in Train.csv)

4. Data Preparation & Cleaning

- **Missing Values:**
  - `Item_Weight` imputed with item-wise mean; global mean used as fallback.
  - `Outlet_Size` imputed using the most common size for each outlet type.
- **Categorical Standardization:**

- Unified inconsistent labels in `Item_Fat_Content`.

- **Duplicates:** Removed any duplicate entries.
- **Outlier Handling:**
  - `Item_Outlet_Sales` values capped at the 99th percentile (Train.csv only).
- **Test Data:** Cleaned without attempting to process (or reference) the absent target column.

5. Feature Engineering & Encoding

- **Transformation Techniques:**
  - Used label encoding for product- and outlet-related categorical variables (to match model training).
  - Created derived features: e.g., outlet age (current year minus establishment year).
- **Feature Set Used for Modeling:**
  - All cleaned and encoded columns with strong business or statistical relevance, including both product and outlet descriptors.

6. Modeling

- **Model Selected:**
  - **Random Forest Regressor:** Chosen for its robustness with categorical and numerical data, ability to model non-linearity, and feature importance analysis.
- **Model Training:**
  - 80/20 train-test split done randomly (due to lack of time series granularity).
  - Ensured no data leakage; model fitted only on training data.
- **Hyperparameters:**
  - Grid-tuned (450 trees, max depth 18) based on holdout validation.
- **Results:**
  - **RMSE**: 1,050.97
  - **MAPE**: 57.89%
  - **R^2 Score**: 0.58

    - The model captures 58% of sales variance, a solid baseline consistent with published ML solutions on similar data.

7. Evaluation & Visualizations

- **Actual vs. Predicted Sales Plot:**
    - Line chart demonstrates that the model follows sales trends, with most predictions close to actuals but visible deviations at sales spikes.
- **Feature Importance Plot:**
    - Bar chart clearly shows `Item_MRP`, `Outlet_Type`, and `Outlet_Location_Type` as top predictors of sales.
- **Residual Analysis Plot:**
    - Scatterplot of residuals vs. actual sales reveals most errors are random; some higher error at high sales magnitudes.
- **Forecasted Sales Trend Plot:**
    - Smoothed line with confidence bands illustrates model-extrapolated sales trends, supporting strategic outlooks.

8. Business Insights & Strategic Recommendations

- **Core Insights:**
    - Pricing (MRP), outlet type, and outlet location are primary drivers of sales performance.
    - Inventory can be optimized by focusing on high-sales categories (Food items, Tier 1 outlet locations).
    - Significant sales peaks align with previously known promotional/seasonal effects.
- **Recommendations:**
    - Stock high-performing items before forecasted peaks.
    - Consider targeted promotions for underperforming outlets/categories.
    - Dynamic pricing strategies could yield higher revenue, given the sales-MRP correlation.
    - Invest in data collection: transaction dates and promotions would enable even stronger models.

9. Limitations & Next Steps

- **Data Limitations:** Lack of time-based transaction data limits time-series approaches.
- **Modeling Constraints:** Random Forest is robust, but ensemble boosting (e.g., XGBoost, LightGBM) may yield further gains.
- **Interpretability:** Random Forest allows basic feature importance, but advanced SHAP or LIME explanations could be used for deeper transparency.

10. Summary Table of Key Model Metrics

Model	RMSE	MAPE	R^2 Score
Mean Baseline	2750	45%	0.00
Random Forest	1051	57.89%	0.58

11. Project Workflow Decisions

- **Used only training data for model building, EDA, and validation.**
- **Used cleaned, encoded Test.csv for prediction and submission only.**
- **All visualizations, except sample feature profiles, were based exclusively on training data (sales must be present).**

12. Resulting Files (for Submission/Reporting)

- `cleaned_bigmart_data.csv` — Clean, encoded training data.
- `bigmart_rf_model.pkl` — Trained Random Forest model with encoders.
- `model_performance_comparison.png` — Model metrics plot.
- `actual_vs_predicted_sales.png` — Line chart for Slide 8.
- `feature_importance.png` — Feature importance plot.
- `forecasted_sales_trend.png` — Forecasted trend with CI.
- `residual_analysis_plot.png` — Residual analysis scatter plot.
- `bigmart_test_predictions.csv` — Predicted sales for Test.csv (Kaggle submission format).

13. Conclusion

This project established a transparent, reproducible sales forecasting pipeline using real retail data. All design choices—from imputation and encoding through to model selection and evaluation—were grounded in best practices for explainable, business-relevant machine learning. The full workflow is ready for future improvements, deployment, and integration with broader business analytics.

Prepared by: Mayank Lalwani
Date: 31/07/2025

⁂