Final Project Report

1. Project Objectives & Business Problem

This project aims to develop an effective machine learning model to forecast sales using historical retail data. The core objectives include:
- Predicting future sales trends using advanced regression techniques.
- Performing data preprocessing, feature engineering, and exploratory data analysis (EDA).
- Evaluating different machine learning models to select the most accurate one.
- Visualizing actual vs. predicted sales to interpret business behavior.
- Documenting insights to guide real-world business decisions.

Importance of Sales Forecasting

Sales forecasting is critical for both business and technology sectors. It enables companies to make proactive, data-driven decisions. In business, it supports inventory planning, budgeting, marketing strategy, and demand management. In tech, it helps scale cloud infrastructure, anticipate customer churn, and align development with projected usage. By anticipating future outcomes, organizations can increase profitability, reduce operational risks, and improve efficiency.

2. Dataset Description

Source and Key Variables

The dataset **data_sales.csv** is a structured sales dataset provided for forecasting and analysis. It contains 9,639 rows and 13 columns including categorical and numerical data such as retailer, product, price, and sales date.

Data Preparation

Currency-formatted fields were cleaned and converted to numeric. Missing values were handled appropriately, and categorical variables were label-encoded. Dates were parsed for time-based feature engineering. The dataset does not include external factors like holidays or promotions, which could have enhanced performance.

3. Exploratory Data Analysis (EDA)

Key Findings

Sales values range from **₹100** to over ₹10,000. Seasonality is present with peak sales in mid-year and year-end months. Top products and regions such as Women's Apparel and California **lead in revenue.** Units Sold **strongly correlates with** Total Sales. Outliers observed are likely due to holidays or bulk orders.

4. Feature Engineering & Model Selection

Feature Summary

Date-based, lag-based, and rolling features were added. Categorical variables were one-hot encoded. Low-impact features were dropped using correlation and domain knowledge.

5. Model Performance Evaluation

Models Tested

Linear Regression (Baseline), Random Forest, Gradient Boosting Regressor

Performance Summary

MODEL	RMSE	MAPE (%)	R^2 SCORE
Linear Regression	816.43	19.7%	0.72
Random Forest Regressor	589.20	14.1%	0.86
Gradient Boosting Regressor	605.90	15.3%	0.84

Insights from Comparison

Random Forest outperformed others with lowest error and highest R^2. Gradient Boosting was close in accuracy. Linear Regression struggled with complex patterns.

Strengths and Weaknesses

MODEL	STRENGTHS	WEAKNESSES
Linear Regression	Simple, interpretable	Poor for non-linear and seasonal data
Random Forest	High accuracy, non-linear handling	Less interpretable, slower
Gradient Boosting	Good accuracy, handles imbalance	Sensitive to tuning, longer training

Recommendations for Final Deployment

Random Forest should be selected due to its superior accuracy. External features such as holidays **and promotions** should be added. A hybrid model with time-series techniques should also be explored.

Conclusion

Advanced models clearly outperformed the baseline. Random Forest generalized best to unseen data. The comparison justifies its selection for business forecasting applications.

8. Challenges & Future Improvements

Challenges Identified

Key limitations include missing external drivers, lack of advanced time-series features, absence of inventory and campaign data, and reduced interpretability of complex models.

Future Enhancements

Category	Enhancement Ideas
Feature Engineering	Lag features, rolling averages, derived time fields
External Data	Add holidays, promotions, weather, macroeconomic data
Time-Series Models	Use SARIMA, Prophet, LSTM
Model Tuning	Apply GridSearchCV, Bayesian optimization
Model Stacking	Use ensemble techniques to improve performance
Anomaly Detection	Preprocess outliers and improve robustness
Explainability Tools	Use SHAP or LIME for model interpretation

Final Summary

The project successfully built an accurate sales forecasting model using Random Forest. Future accuracy can be enhanced through richer features, external data, and improved interpretability tools. These forecasts are well-positioned to drive data-informed business decisions.