**Design Document: Bakery Sales Prediction Using Machine Learning**

**1. Problem Statement**

We are given a dataset with daily sales of a bakery. The goal is to create a machine learning model that can predict the number of sales for the next N days.

**2. Dataset Overview**

* **File**: dataset.csv
* **Columns**:
  + DATE: Date of the sale (MM/DD/YY format)
  + SALES: Number of items sold on that date

**3. Approach**

**Step 1: Data Preprocessing**

* Convert DATE column to datetime format
* Sort the data by date
* Create a new feature Days representing the number of days since the first date
* Create a new features like dayofweek,month,day,is\_weekend.

**Step 2: Model Selection**

Two models were trained:

1. **Linear Regression**
2. **Random Forest Regressor**

**Step 3: Evaluation Metrics**

* **MSE ( Mean Squared Error) and MAE**: Lower is better
* **R² Score**: Closer to 1 is better

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R² Score** |
| Linear Regression | 1012.14 | 0.20 |
| Random Forest Regressor | 852.51 | 0.33 |

**Step 4: Model Selection Justification**

**Random Forest Regressor** was selected as it had a lower MSE and higher R² Score compared to Linear Regression, indicating better predictive performance.

**Step 5: Future Prediction**

To predict the next N days, we calculate the future day indices (as Days since start), and use the trained Random Forest model to predict the sales.

**4. Libraries Used**

* pandas
* numpy
* sklearn (for modeling)
* matplotlib and seaborn ( for plots)

**5. Output Format**

The final script produces predictions in the following format:

DATE SALES

0 YYYY-MM-DD NN

1 YYYY-MM-DD NN

...

**6. Script to run:**

Run the prediction using:

python predict.py N

Where N is the number of future days to predict.

**7. Notes**

* Model is trained on historical trend without external factors.
* Can be further improved by adding features like holidays, promotions, etc and using different ML Models.