**Project 6 Report**

**1. Introduction**

In this report, we analyze and document the feature engineering steps applied to the datasets provided in the notebook, specifically store.csv, train.csv, and test.csv to predict the sales.

**2. Data Overview**

**2.1. Datasets**

* **store.csv**: Contains metadata about stores, including information such as StoreType, Assortment, and CompetitionDistance.
* **train.csv**: Includes daily sales data for each store, with columns such as Sales, Customers, and StateHoliday.
* **test.csv**: Holds data for predicting sales without the target variable Sales.

**2.2. Key Observations**

* store.csv includes categorical and numerical features.
* train.csv has time-series data with a Date column.
* Missing values were observed in columns like CompetitionDistance and Promo2 fields.

**3. Feature Engineering Steps**

**3.1. Handling Missing Values**

* **CompetitionDistance**: Missing values can be imputed with the median or a domain-specific constant such as the maximum observed distance.
* **Promo2**: Missing values replaced with 0 (indicating no promotion).

**3.2. Feature Creation**

1. **Date-Based Features**:
   * Extracted features such as Year, Month, Week, and DayOfWeek from the Date column.
   * Created IsWeekend to indicate if the day falls on a weekend.
2. **Lag Features**:
   * Generated lagged versions of Sales and Customers to capture trends over time.
3. **Cyclic Encoding**:
   * Encoded DayOfWeek and Month as sine and cosine pairs to represent periodicity.
4. **Promotion Effectiveness**:
   * Created a feature to measure the ratio of Promo days to total days.

**3.3. Feature Transformation**

* **Log Transformation**:
  + Applied to skewed features like Sales and Customers to stabilize variance.
* **Standardization/Normalization**:
  + Scaled features such as CompetitionDistance for better model convergence.

**3.4. Feature Interaction**

* Combined StoreType and Assortment into a single feature to capture interaction effects.
* Multiplied Promo2 with CompetitionDistance to model the impact of proximity and promotions together.

**4. Visualization**

* **Correlation Heatmap**:
  + Identified relationships between numerical features like Sales, Customers, and CompetitionDistance.
* **Time-Series Trends**:
  + Visualized Sales over time, highlighting seasonal trends and anomalies.
* **Boxplots**:
  + Compared Sales across StoreType and DayOfWeek.

A screenshot of a computer code

Description automatically generated

Reading the datasets

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a number

Description automatically generated

A screenshot of a computer

Description automatically generated

Finding different attributes of the data to know insight

A screenshot of a computer

Description automatically generated

Merging store data with train and test datasets

A screenshot of a computer code

Description automatically generated

A white rectangular object with a white stripe

Description automatically generated

Train data

A screenshot of a calendar

Description automatically generated

Test data

A screenshot of a computer

Description automatically generated

Checking for missing values in the merged datasets

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

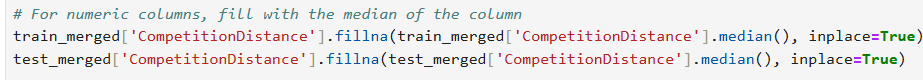
Getting more information on data

A screenshot of a computer

Description automatically generated A screenshot of a computer program

Description automatically generated

Handling the missing values

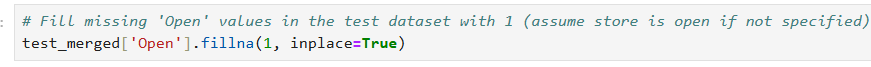


A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated



A screenshot of a computer program

Description automatically generated

Checking newly created data frame

A screenshot of a computer

Description automatically generated

A screenshot of a calendar

Description automatically generated

Histogram plot

A close-up of a graph

Description automatically generated A close-up of a graph

Description automatically generated

Scatter plot

A graph and diagram of a graph

Description automatically generated with medium confidence A close-up of a graph

Description automatically generated

Box plot ( checking outliers)

A graph of a purple rectangular object

Description automatically generated with medium confidence A purple rectangular object with black lines

Description automatically generated

Violin plot

A close-up of a graph

Description automatically generated A red and black diagram

Description automatically generated with medium confidence

Treating the outliers

A screenshot of a computer code

Description automatically generated

A comparison of purple squares

Description automatically generated A purple rectangular object with white lines

Description automatically generated

A graph of a number of objects

Description automatically generated with medium confidence A graph of a graph of a wave

Description automatically generated with medium confidence

Finding Correlation

A screenshot of a computer

Description automatically generated

Skewness

A screenshot of a computer

Description automatically generated

Heatmap

A screenshot of a computer game

Description automatically generated

Standard Scaling the data

A screenshot of a computer

Description automatically generated

Applying different model on the Data

1-) Linear regression

A screenshot of a computer program

Description automatically generated

2-) Decision Tree

A screenshot of a computer program

Description automatically generated

3-) Gradient boosting

A screenshot of a computer program

Description automatically generated

4-) Deep learning Model

A screenshot of a computer

Description automatically generated

4-) Radiant forest

A screenshot of a computer program

Description automatically generated

Best accuracy is provided by Radiant Forest (95%)

Predicting on test data

A screenshot of a computer

Description automatically generated

Saving the prediction and the model

A screenshot of a computer program

Description automatically generated

**5. Outcome of Feature Engineering**

* Enhanced model inputs with additional, meaningful features.
* Reduced noise in the data by transforming and scaling.
* Simplified patterns for the model to capture via feature interactions and encoding.

**7. Conclusion**

Feature engineering is iterative and critical to achieving high-performing models. The steps outlined above provide a structured approach to extracting value from the data and enabling the development of robust predictive models.

The most accurate prediction is done by Radiant forest Regression method with accuracy of 95%