Final Project Report

Project Overview

This project aims to predict the best **Location** and **Holiday** setting that maximizes profit for a business. We designed and trained machine learning models based on a generated synthetic dataset that includes various features like weather, population density, season, and competitor data.

- **Models Implemented:**
  - Linear Regression
  - Random Forest Regressor
- **Files Generated:**
  - `performance_summary.txt`: Contains model performance evaluation (MSE and R2 scores) with timestamps.
  - `results_summary.txt`: Contains best location and holiday predictions at each model run, labeled with timestamps.

How the Code Works

1. **Dataset:**
   - Reads a synthetic dataset file named `generated_profit_data.csv`.
2. **Preprocessing:**
   - Categorical variables are encoded using One-Hot Encoding.
   - Numerical variables are scaled using Standard Scaler.
   - PCA is applied to retain 95% variance.
3. **Model Training:**
   - Data is split into training and testing sets.
   - Two models (Linear Regression and Random Forest) are trained.
   - Model performance (MSE and R2) is printed and saved.
4. **Prediction:**
   - The model predicts the best location and holiday combination by simulating possibilities.
   - Results are printed to console and saved to a text file.
5. **Outputs:**
   - All results are appended into output files with a timestamp for easy tracking across multiple runs.

Execution Instructions

1. Install the required libraries:

```
pip install pandas numpy scikit-learn
```

2. Ensure the following structure:

```
FinalProject/
├── models/
│   └── predict_best_location_holiday.py
├── generated_profit_data.csv
```

3. Run the script:

```
cd models/
python predict_best_location_holiday.py
```

4. Review Outputs:
    - `performance_summary.txt`: Model performance metrics.
    - `results_summary.txt`: Best predicted Location and Holiday.

Special Notes

- Each model evaluation and prediction is clearly separated with timestamps.
- PCA helps reduce complexity without losing too much information.
- Random Forest is typically more accurate than Linear Regression for this dataset.